NTNU – Trondheim
Norwegian University of
Science and Technology

Sentiment Analysis for Financial Applications

Combining Machine Learning, Computational Linguistics, and Statistical Methods for Predicting Stock Price Behavior

Lars Smørås Høysæter
Pål-Christian S Njølstad

Master of Science in Computer Science
Submission date: June 2014
Supervisor: Jon Atle Gulla, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Abstract

In this thesis we use sentiment analysis, a classification task within the field of artificial intelligence, for financial applications. Hereunder, we combine machine learning, computational linguistics, and statistical methods for anticipating stock price behavior of ten shares listed on the Oslo Stock Exchange (OSE). These predictions have been made on the basis of sentiment classifications of firm-specific news articles, output by our specially constructed sentiment engine, and an aggregated market-wide sentiment index. The motivation for this approach comes from news being a most felicitous source of financial information; in effect a widely-read filtering and aggregating funnel of sentiments. Furthermore, the OSE has been selected, firstly, for its faculty of being inefficient, compared to peer marketplaces, and, secondly, for the inherent barriers to processing the Norwegian language associated with the exchange, having meagre linguistic resources. If able to surmount these barriers and exploit the predictive value of news sentiments, one could potentially attain a competitive advantage trading in this market.

In constructing the named sentiment engine, we have found contextual features to be paramount in classification precision in addition to having developed and optimized a parsimonious approach to sentiment lexica construction. Despite the lack of linguistic resources, we achieve state-of-the-art classification precision in this approach using manual annotation. The engine has been found to make statistically significant predictions on stock return, volume, and order size. Positive articles, predominantly, lead to significant increases in volume while negative articles predict the opposite effect. The same is the general proclivity for order size. For return, only negative articles impact future stock price behavior, ceteris paribus, depreciating subsequent stock prices. The interaction between news articles and market-wide sentiment is also statistically significant. Although the sign of this latter effect seems firm-idiosyncratic, our analysis reveal that illiquid stocks exhibit stronger reactions than liquid stocks.

I denne oppgaven bruker vi sentimentanalyse, en klassifiseringsoppgave innen fagområdet kunstig intelligens, for finansielle anvendelser. Herunder kombinerer vi maskinlæring, datalingvistikk og statistiske metoder for å forutse kursutviklingen til ti aksjer notert på Oslo Børs (OSE). Disse prediksjonene er gjort på grunnlag av en sentimentklassifisering av selskapsspesifikke nyhetsartikler, generert av vårt spesialkonstruerte sentimentanalysesystem, og en aggregert, markedsgenerell sentimentindeks. Motivasjonen for denne tilnærmingen er at nyheter kan være en særs god kilde til finansiell informasjon; allment lest og med evnen til å filtrere samt aggregere meningsytringer fra flere kilder. Videre har Oslo Børs blitt valgt, for det første, for sin mangel på effesiens, sammenlignet med andre markedsplasser, og for det andre, for hindringene knyttet til at børsens nyhetsstrøm er skrevet på norsk – et språk med mangel på lingvistiske ressurser. Hvis man er i stand til å overvinne disse hindringene, samt utnytte den prediktive verdien av nyhetssentiment, kan man potensielt oppnå et konkurransefortrinn ved handel i nettopp dette markedet.

Ved utarbeidelsen av sentimenanalysesystemet har vi avdekket at kontekstuelle attributter er avgjørende for klassifiseringspresisjon. I tillegg har vi utviklet, og optimalisert, en sparsommelig tilnærming til sentimentleksikonkonstruksjon. Til tross for mangelen på språklige ressurser, oppnår vi state-of-the-art presisjon i klassifisering ved bruk av manuell annotering. Systemet har vist seg å kunne forutsi avkastning, volum og ordrestørrelse på statistisk signifikant vis. Positive artikler fører til betydelig økning i volum, mens negative artikler gir motsatt effekt. Det samme er den generelle tilbøyeligheten til ordrestørrelse. For avkastning vil bare negative artikler påvirke den fremtidige kursutviklingen og vil, alt annet ved like, gi en reduksjon av denne. Samspillet mellom nyhetsartikler og markedsgenerelt sentiment er også funnet statistisk signifikant. Selv om fortegnet til denne siste sammenhengen ser ut til å være selskaps-spesifikt bestemt, viser våre analyser at mindre likvide aksjer reagerer sterkere enn de mer likvide.

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU), as partial fulfillment of the degree of Master of Science (Sivilingeniør), and as part of the course *TDT4900 Computer Science, Master's thesis*. The work culminating in this report has been performed at the Department of Computer and Information Science (IDI), under the supervision of Professor Jon Atle Gulla, and as a part of the SmartMedia project.

(This page is intentionally left blank.)

Acknowledgements

First and foremost, we would like to express our humble gratitude towards our supervisor Professor Jon Atle Gulla for his persistent guidance, indefatigable encouragement and inexorable dedication to the publication of our work. Furthermore, we would like to thank Senior Engineer Arne Dag Fidjestøl at IDI for setting up and maintaining the Oslo-Bergen Part-Of-Speech Tagger that we used in feature extraction. We would like to thank Dr. Jon Espen Ingvaldsen for providing us with the interface to the SmartMedia news database, the basis of our dataset. Additionally, we are grateful for the early dialogues with Professor Lars Hellan and Associate Professor Dorothee Beermann Hellan on the linguistic aspects of our project. We would also like to thank Assistant Professor Bei Yu of Syracuse University (USA) for giving us input on which news categories to focus on in our analyses and sharing her experience on annotation studies. We are also indebted to Dr. Wei Wei for his helpful comments on the first paper presented in this thesis, where he is a co-author. Lastly, we are grateful for Associate Professor Øyvind Salvesen, co-author on our third paper, helping us find the appropriate statistical methods to use for the analysis of our financial dataset. Nonetheless, all errors are our own.

(This page is intentionally left blank.)

Contents

List of Figures

(This page is intentionally left blank.)

List of Tables

(This page is intentionally left blank.)

Part I

Research Overview and Summary

(This page is intentionally left blank.)

Chapter 1

Introduction

In this chapter we introduce the research conducted within the scope of this thesis. In Section 1.1 the background and motivation of the work is presented. An outline of the problems addressed in the thesis are provided in Section 1.2. In section 1.3, the research goals and questions are listed and elaborated on before we account for the contributions made in Section 1.4. Three papers that have been written as a part of this thesis. How these papers are related, along with high-level descriptions of each of them, is explained in Section 1.5. Finally, a structured overview of the remainder of the thesis is given in Section 1.6.

1.1 Background and Motivation

With the evolvement of the Internet, the amount of readily available information has grown exponentially [Huberman and Adamic, 1999]. This copious display of available information, generated by a large number of users, represents a valuable source for decision-making. It is, however, far from nontrivial to systematically make sense out of this panoply of information at a large scale. Traditional search technologies, like that of Google Inc., have made, and continues to make, vast improvements in systematically searching explicitly contained information in Internet documents. Analyses seeking to exploit implicitly contained information, like the sentiments expressed by the author of a text, are much less researched.

Presumably, these kinds of analyses could prove just as valuable to certain decision-makers in an increasingly fast-paced and competitive world. In fact, an organization's ability to effectively make sense out of such opinions expressed in Internet documents could very well turn out to be key in maintaining competitiveness and fulfilling its purpose going forward. For instance, a company manufacturing products could keep the opinionated pulse of its existing and prospective

customers as it considers which new products to develop [Yi et al., 2003]. A mutual fund holding a diversified portfolio of stocks could monitor the sentimental landscape associated with its current assets opined by journalists of various financial news papers [Parikh et al., 2012]. A hedge fund, or any other trader of financial instruments or assets, could process sentiments from news stories, filings, social media, and blogs in real-time and make trades in anticipation of future price behaviors [Andrews et al., 2011], which, according to Feldman [2013], can lead to superior returns. This latter application, which in our case will be fueled by sentiments from financial Internet news articles, is exactly the subject of this thesis.

Undeniably, as the number of opinionated documents of interest grows it becomes practically impossible to manually analyze all these sentiments in an effective manner. Hence, research addressing the scalability of this problem has emerged, and soon evolved, drawing on theory from the intersection machine learning, computational linguistics, and, in this specific case, finance [Wei and Gulla, 2010]. Early influential work on sentiment analysis, as this task of automatically classifying the sentiments expressed in document is usually referred to as, includes that of Hu and Liu [2004], Pang and Lee [2004], Pang et al. [2002], Hatzivassiloglou and McKeown [1997], in addition to Dave et al. [2003]. Research in this subfield of artificial intelligence continues to showcase improvements. Recently, Socher et al. [2013] claimed to push the single-sentence state-of-the-art sentiment classification precision from ∼80% to 85.4%. Content-wise, document domains composed of news articles [Balahur et al., 2013], social media text snippets [Pang and Lee, 2008], blog posts [Chesley et al., 2006], and political speeches [Yu et al., 2008], among others, have been subject to extensive research.

Hence, the motivation for this project comes from the demand for, and in extension of recent advances in, adding sentiment analysis to financial Internet news articles before using this to make predictions on stock price behavior. With this, we hope to unveil remunerative causal relationships between the news flow associated with listed stocks and their subsequent price movements. Although sentiments can appear in numerous different forms, and be proxied in equally plentiful ways[1], we argue that the news is the most felicitous source of financial sentiments. The reason for this is fourfold: news are 1) firm-specific in that they easily can be linked to specific stocks (and can equally importantly be deemed non-relevant for other stocks), 2) rich in expressiveness, 3) acting as a funnel of sentiment - filtering and aggregating the reports of multiple different journalists, and 4) widely read by market agents. To illustrate the meaning of news acting as a funnel of sentiments we have visualized a simple model of this in Figure 1.1.

Figure 1.1 shows a simple model with multiple financial markets being covered,

[1]See Baker and Wurgler [2007] for an elaborate list of sentiment proxies having been used in financial market settings.

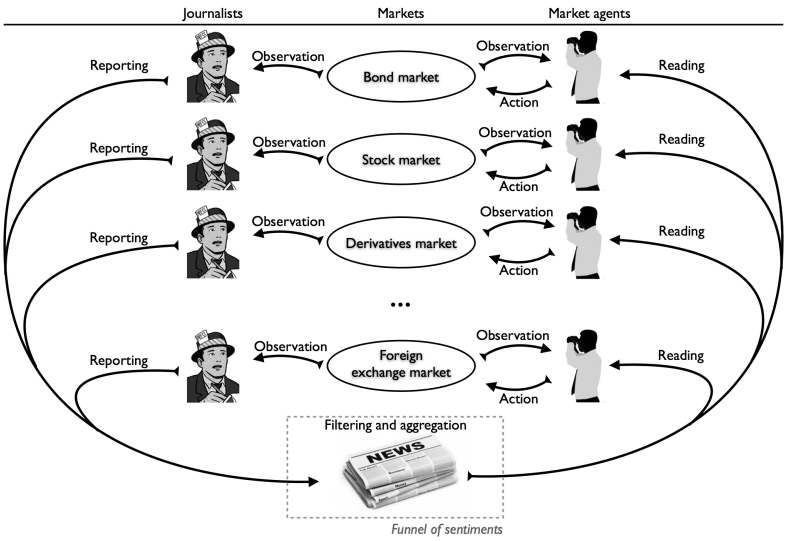


Figure 1.1: Simple model of news as a funnel of sentiments

observed and acted upon by different journalists and market agents. The reports written by journalists are filtered (only the most important ones make it to the print) and aggregated (events in different markets are put in context of each other) by editors before finally being published in the news. These news are then widely read by the market agents. As such, the agents get a much broader view of the overall market sentiments, than what is visible in the specific market(s) that she is actively engaged. In this sense, the news act as a funnel of sentiments and is, in this capacity, a most felicitous source for sentiment analysis.

In our research, we will focus our attention to the news flow of stocks listed on the Oslo Stock Exchange (OSE). The vast majority of this news flow is written in the Norwegian language and, hence, this will be the language of our analysis. The reason for this focus is twofold: 1) as Gjerde and Saettem [1999] establish, the Norwegian financial market is less mature compared to that of the U.S. and other peers, presumably leaving greater opportunities for exploiting sentiment reactions, and 2) the linguistic barriers[2] to developing a system for analyzing the stock price reaction to news are much greater, leaving a preeminent upside if successfully able to do so. The stocks listed on OSE are, when compared to those listed on larger exchanges like New York Stock Exchange (NYSE), generally less

[2]The availability of lexical and linguistic resources is much more limited for Norwegian than for other languages like English.

efficient. According to works on value investing, like that of Graham et al. [1934] and Graham [1959], such less efficient stocks are exactly the ones an 'intelligent investor' should seek to analyze.

1.2 Problem Outline

The main problems studied in this thesis are the identification and quantification of author opinion in financial Internet news articles and, subsequently, the causal linking of these to stock price behavior. The former is known as sentiment analysis and is well researched. More than 7000 articles have been published on the topic, hundreds of startups are developing solutions for sentiment analysis, and major statistical packages such as Statistical Analysis Software (SAS) and Statistical Package for the Social Sciences (SPSS) now include dedicated sentiment analysis modules [Feldman, 2013]. However, both within the domain of financial Internet news articles and for the Norwegian language, which is the subject for our study, this type of analysis is much less studied. Some of the complications and challenges associated with these two domains are detailed in Section 1.2.1 and 1.2.2, respectively. Furthermore, the problem of causally linking sentiments to stock price behavior is discussed in Section 1.2.3.

1.2.1 News Domain

Within sentiment analysis most research has been conducted in texts where there is a clearly defined target and this is unique across the texts, including product reviews, blogs, political speeches, and debates. News articles do not always have a clearly defined target nor are the targets necessarily unique across articles. Furthermore, attributes like the text length (of title, lead and main), language style, quotation usage (implicit and explicit), use of irony or sarcasm varies greatly, not only between different news categories but also between articles within the same category. All of these divergences from the corpora usually used in sentiment analysis makes this kind of analysis much harder within the news domain. For this reason, not all categories of news articles are well-suited for sentiment analysis.

Moreover, sentiment-constituting expressions and words will vary between these categories. For instance, the word 'bull' in a financial news context pertains to a positive market trend and not to the mammal as would be the case in other news domains. Hence, sentiment analysis classifiers would have to be developed for suited domains and these would have to be tailored specifically to efficiently capture the sentiment-bearing expressions and words in this context.

One specific news category that is suited for this type of analysis, as it typically is associated with sentiments, is finance. News in this category often relate events

to market expectations, reflected in the current stock price valuation, or make forward-looking predictions about the future. This is necessarily subjective, and usually sentiment bearing, since no one knows what the market is thinking or what future will hold. Examples of such predictions include analysts claiming that a stock is undervalued, that the interest rate will be raised by the central bank, or that a firm is deemed likely to win a major contract.

1.2.2 Language of Analysis

Nearly all current sentiment analysis systems are constructed to work only in a single language, usually English. Effective sentiment analysis requires thorough understanding of the language that the opinions are expressed in [Bautin et al., 2008]. To our knowledge, little to no research has been conducted within sentiment analysis of the Norwegian language. There have, however, been some commercial attempts by the Norwegian software company Cxense[3] to perform sentiment analysis for the online news paper Adressa[4]. The results of these attempts are, albeit, not publicly available. In lack of previous work, there are no available resources for conducting such analyses, as is the case for the English language. Examples of such resources are lexica, with sentiment values for word entries, data structures for evaluation of the sentiment values of newly encountered words, like word graphs, among others. Furthermore, the linguistic constituents defining sentiments in the Norwegian language are, not only less understood, but undescribed. For instance, research in the English language has resulted in the identification of Valence shifter[5] lists, like it has been done in Polanyi and Zaenen [2006] and Simančík and Lee [2009]. There exists no equivalent for the Norwegian language. Although most sentiment analysis systems are language specific, the techniques and principles could, with some effort, be reapplied to Norwegian. This reapplication is, however, complicated by the lack of available linguistic resources which will resultin in insensible, high levels of effort [Bautin et al., 2008].

Independent of application domain, sentiment lexica usually play a central role in performing the required classification task [Lu et al., 2011] and the consensus within the research field of sentiment analysis is that domain-independent universal sentiment lexica are futile [Qiu et al., 2009; Lu et al., 2011; Turney, 2002]. Wilson et al. [2005] quantify this inferiority; they found a universal lexicon to yield a precision of merely 48% whereas a contextual devised lexicon achieved 65.7%. As such, much effort has been directed towards building domain-specific

[3] www.cxense.com

[4] www.adressa.no

[5] Words or phrases that, in certain contexts, change the sentiment polarity of a phrase or sentence. Examples in the English language include *not*, *never*, and *nor*.

lexica. The two most popular approaches to doing so are 1) automatically extracting lexicon entries, with classification, from a collection in a new domain [Godbole et al., 2007; Tan et al., 2008; Lu et al., 2011] and 2) automatically extending sentiment lexica from a known domain or application to others [Qiu et al., 2009; Neviarouskaya et al., 2009]. These approaches, however, require having either a fairly large dataset, in order to use statistical methods for entry extraction, or the availability of ample lexical resources, like a sentiment lexicon or a synonym dictionary. Within some application domains, neither of these requirements can easily be met. In our specific case, the corpus size is small since we perform sentiment analysis of the news flow associated with stocks listed on the (less covered) OSE exchange over a limited period of time. Additionally, since the associated news flow is in Norwegian, publicly available sentiment lexica and synonym dictionaries are not available for our purposes, as previously noted [Perez-Rosas et al., 2012].

Hence, for our domain, neither automatic sentiment lexicon with classification extraction nor lexicon extension are viable approaches. In lieu of these techniques, acquisition of a sentiment lexicon in a supervised fashion represents a feasible option. What then needs to be investigated is 1) how much effort this requires and 2) how to best carry out this procedure whilst achieving satisfactory sentiment classification performance.

1.2.3 Predicting Stock Price Behavior

Regression analysis is the statistical process typically employed for estimating the relationship between variables - which in our case is the relationship between news flow and stock price behavior. Specifically, the Ordinary Least Squares (OLS) linear regression model is popular in econometric analyses, which seems imminent for this study, due to its desirable property of being the best linear unbiased estimator (BLUE) of coefficients in a linear regression model. However, this property only holds under certain assumptions, which in our case cannot be satisfied since the stock price behavior of shares listed on the OSE, like most other shares, exhibit time-varying volatility. Naturally, some periods, like during the nadir of the financial crisis, display far more stock price fluctuations than others. This is a common problem encountered when trying to predict financial variables [Campbell and Andrew, 1997] and can be mitigated by finding other, more sophisticated, regression models suited for the evident heteroskedasticity [Wooldridge, 2012]. Under certain assumptions, the autoregressive conditional heteroskedasticity (ARCH) model can be used [Engle, 1982]. This means that lagged error terms, able to adjust for this time-varying volatility, are added to the regression equation. Furthermore, it is unreasonable to assume the variance of two completely different stocks to be the same and, thus, separate ARCH

regression models need to be fit for each of the stocks considered. Moreover, this can make it harder to make inference on the relationship between variables across regression models. For this purpose, Wald tests, or other parametric statistical tests, need to be put into use.

1.3 Research Goals and Questions

The research in this thesis aims to develop a system for sentiment analysis of Norwegian financial Internet news articles for prediction of stock price behavior. This necessitates 1) the construction of a sentiment engine to be used for classification of the Internet news articles in question and 2) the evaluation of this engine on the news flow associated with stocks listed on the OSE. Additionally, we will, 3) be studying the interaction impact of firm-specific and aggregate market-wide sentiment on stock return, volume and order size. The research goals and the necessary steps to achieve these goals, in the form of research questions, will now be briefly introduced and listed in turn.

1.3.1 Sentiment Engine Construction

In order to create a sentiment engine achieving satisfactory classification precision, the two challenges outlined in the previous section need to be addressed: we need to tailor sentiment analysis for the financial domain, which we will do by investigating which feature categories are most important in classification (RQ1), in addition solving the problem of the limiting availability of lexical and linguistic resources for the Norwegian language. This we will do by optimizing a parsimonious method for manually constructing sentiment lexica (RQ2).

RQ1 Which feature categories, input to machine learning methods, are the most important when performing sentiment analysis on financial Internet news articles?

RQ2 How can sentiment lexica for a language with meager lexical and linguistic resources (like the Norwegian) be devised through manual annotation at permissible levels of efforts?

1.3.2 Sentiment Engine Evaluation

Having devised a sentiment engine for classification of financial Internet news articles, achieving state-of-the-art precision, the ultimate test to this system is to test whether this can make predictions on stock price behavior.

RQ3 Can such a sentiment engine, classifying financial Internet news articles written in Norwegian with state-of-the-art performance, make predictions on stock price behavior?

1.3.3 Interaction impact of firm-specific and aggregate market-wide sentiment on stock price behavior

With sentiments being the topic of this thesis and a classification engine in place, we want to take our analysis further by examining firm-specific and aggregate market-wide sentiments to see if the interaction between these two variables have any predictive value on stock price behavior.

RQ4 Which relationships exist between the interaction of firm-specific news, aggregate market-wide sentiment and stock price behavior?

1.4 Research Contributions

This thesis has four main contributions and these are, in answer to the aforementioned research questions, as follows:

C1 Contextual feature categories are paramount in performing sentiment classification of financial Internet news articles written in the Norwegian language.

As detailed in Paper I (Chapter 6), the contextual feature category proved paramount in performing sentiment classification of financial Internet news articles written in the Norwegian language. Hence, the primacy of finding methods for devising sentiment lexica for our domain was uncovered.

C2 Sentiment lexica for a language with meagre lexical and linguistic resources (like the Norwegian) can be devised through manual annotation at permissible levels of efforts. In this parsimonious approach, COT radius should be kept low, the ranking function mutual information should be used, the lexicon size should be ∼30% of the COT candidate list and the machine learning classifier J48 should be employed to achieve the best precision.

Sentiment lexica, as delineated in Paper II (Chapter 7), for a language with meager lexical and linguistic resources can be devised through manual annotation at permissible levels of efforts, all the while achieving state-of-the-art classification precision. The optimized parameters were used as input to the final sentiment engine used to classify the news flow associated with ten stocks listed of the OSE.

C3 A sentiment engine, classifying financial Internet news articles written in Norwegian with state-of-the-art performance, can be made to reveal predictions on the stock price behaviors of volume and order size. Positive articles predominantly lead to significant increases in volume while negative articles have the opposite effect. The same appears to be the general proclivity for order size. For return, only negative articles are found to significantly impact future stock price behavior and the publication of such articles are largely found to, ceteris paribus, reduce subsequent returns.

C4 The interaction between news articles and market mood is statistically significant. Although the sign of this effect seems firm-idiosyncratic, our analysis revealed that white chips'[6] reactions are of greater magnitude than that of blue chips[7].

As described in Paper III (Chapter 8), applying the constructed sentiment engine on the news flow associated with stocks listed on the OSE revealed several statistically significant relationships between publication count by polarity and traded volume, in addition to average order size. The interaction impact of news articles and market mood on these same dependent variables were also found to be statistically significant.

1.5 Papers

There are three papers included in this thesis. The first (Paper I below) has been accepted for publication in the proceedings of the 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2014). The second (Paper II) has been submitted to Springer's Language Resources and Evaluation and is awaiting notification of acceptance. The third (Paper III) is still a working paper and is planned to be submitted to a relevant journal in empirical or quantitative finance. Details of these papers are listed below and can be found in completeness in Part II of this thesis.

Paper I Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, Wei Wei, and Jon Atle Gulla: *Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News*, to appear in the proceedings of the 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2014).

Paper II Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, and Jon

[6] Stocks with relatively low market value and liquidity.
[7] Stocks with relatively high market value and liquidity.

Atle Gulla: *Optimizing Supervised Sentiment Lexicon Acquisition: Selecting Co-Occurring Terms to Annotate for Sentiment Analysis of Financial News*, submitted to Springer's Language Resources and Evaluation.

Paper III Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, Øyvind O. Salvesen, and Jon Atle Gulla: *The interaction impact of firm-specific news and market wide sentiment on stock price behavior - evidence from the Oslo Stock Exchange (OSE)*, working paper planned submitted to a relevant journal in empirical or quantitative finance.

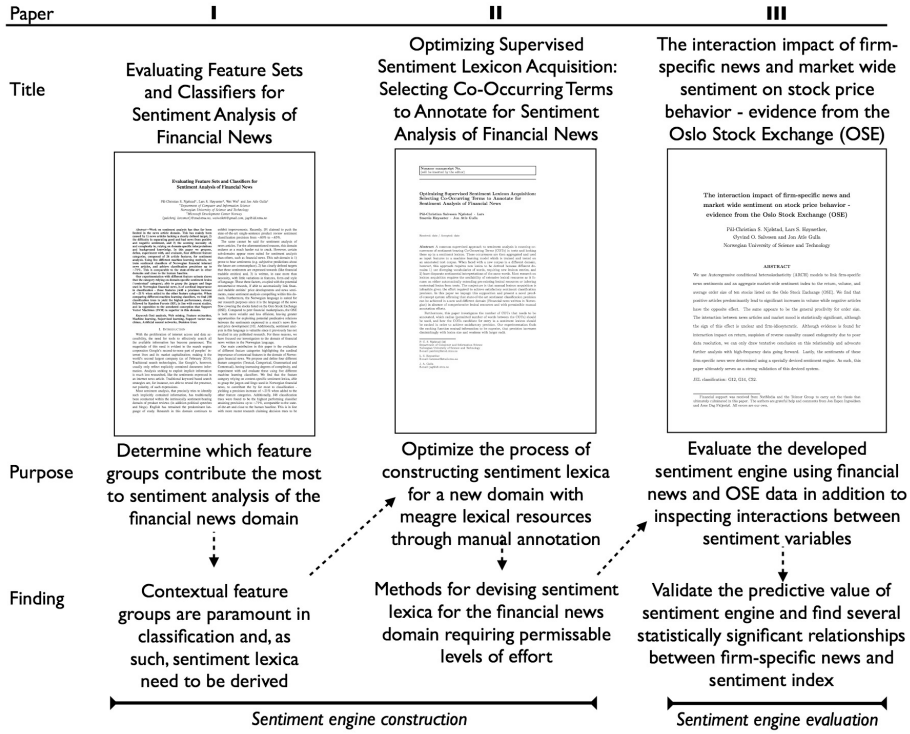


Figure 1.2: Visualization of how the three papers tie together in this thesis.

We will now briefly adumbrate how these papers are related, before giving short, high-level descriptions and accounts for the emphasis for each of them. Since this thesis, as its subtitle clearly states, combines machine learning, computational linguistics, and statistical methods to make financial predictions, its entirety might not be of interest to all readers. Hence, we kindly ask readers only in-

terested in a specific subtopic to skip to the relevant paper in Part II, as these are self-contained contingent familiarity with the general paper topic. Figure 1.2 visualizes how the three papers in this thesis are related. The two first (Paper I and II) document the methods used in constructing the sentiment engine that is then evaluated in paper III. This latter paper also studies the interaction impact of firm-specific news and aggregate market-wide sentiment on stock price behavior. Furthermore, the finding of paper I dictates the purpose of paper II whose finding, constitutes the premise for paper III, as sought illustrated in the same figure.

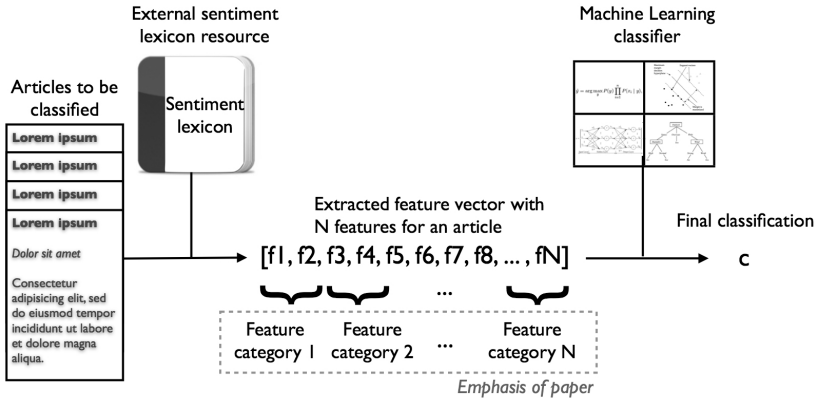


Figure 1.3: High-level description and emphasis of paper I, *Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News.*

In the first paper (Paper I), we study the importance of different feature categories which are input to machine learning methods used for performing sentiment analysis on financial Internet news articles. At a high level, as depicted in Figure 1.3, this is done by first converting each article to be classified into a numerical vector ($[f1, f2, f3, f4, f5, f6, f7, f8, ..., fN]$ in the figure), by the means of an external sentiment lexicon resource. A machine learning model then learns the best-fit function that maps from this input vector to output classification (c in the figure) based on the supplied training and test datasets. In this paper, as indicated by figure's grey dotted box, we specifically investigate how different feature categories (Textual, Categorical, Grammatical and Contextual to be precise) contribute to classification precision. We find the latter Contextual feature category to be paramount in classification.

The second paper (Paper II) acknowledges the finding from Paper I, that contextual features are imperative to sentiment analysis in the financial domain, and investigates how a sentiment lexicon for a language with meagre lexical and

linguistic resources (like the Norwegian) can be devised through manual annotation. In this parsimonious approach, we optimize a number of parameters in this lexicon construction: 1) which radius should be used when extracting Co-Occurring Terms (COTs) from an article collection, 2) which ranking function should be used to order the candidates for entry in the lexicon, 3) how many such candidates should be manually annotated and stored in the lexcion, and, lastly, 4) which machine learning classifier should be used in the final classification. The high-level emphasis of this paper, which all has to do with the selection of COTs to be entered into the external sentiment lexicon resource, is enclosed by the grey dotted box in the figure. In addition to determining the optimal values for the named parameters for analysis of financial news articles, we establish that the creation of a sentiment lexicon can, in absence of ample resources, be created at permissable levels of effort for an entirely new domain. Paper I and II collectively document the methods used in constructing the sentiment engine that subsequently is put to practice in the last paper.

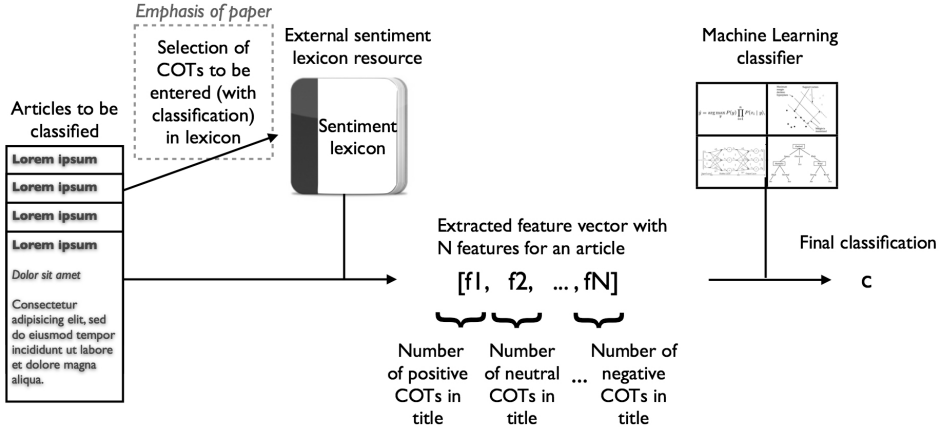


Figure 1.4: High-level description and emphasis of paper II, *Optimizing Supervised Sentiment Lexicon Acquisition: Selecting Co-Occurring Terms to Annotate for Sentiment Analysis of Financial News.*

In the third and final paper (Paper III), we extract the news flow associated with ten articles listed on the Oslo Stock Exchange (OSE), use the devised sentiment engine to classify these as positive, neutral, or negative and, lastly, use these to predict stock price behavior. This is illustrated in Figure 1.5. Specifically, we try to anticipate return, volume and order size of the stocks in question. This paper not only confirms the validity of our developed engine, but also unveils

statistically significant relationships between the interaction of firm-specific news and aggregate market-wide sentiment and stock price behavior.

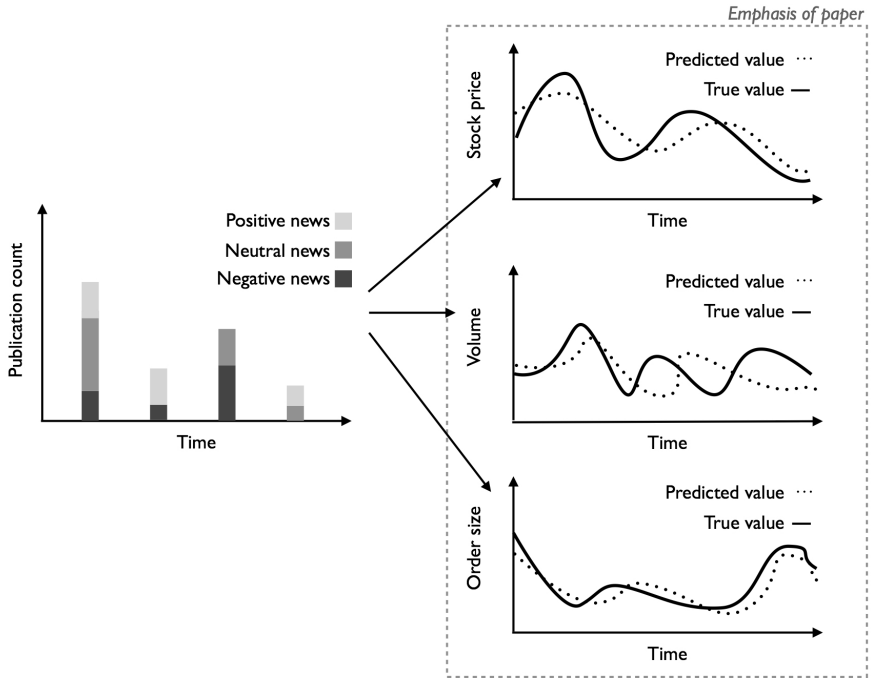


Figure 1.5: High-level description and emphasis of paper III, *The interaction impact of firm-specific news and market wide sentiment on stock price behavior - evidence from the Oslo Stock Exchange (OSE).*

1.6 Thesis Structure

This thesis is composed of two parts. Part I first introduces the background and motivation for the research of our thesis. Furthermore, in this part we account for the technological background relevant for our work(Chapter 2), we briefly review related work (Chapter 3), present results and summarize the contributions of the thesis (Chapter 4) before we conclude and discuss some interesting venues for further work (Chapter 5). Part II includes the three selected papers, in completeness, as detailed in section 1.5.

Part I - Research Overview and Summary

Chapter 1 - Introduction This chapter explains the background and motivation for this thesis in addition to describing research goals, questions, contributions, and outlining the selected papers.

Chapter 2 - Theoretical Background This chapter contains a theoretical overview of the methods used in this thesis. This involves theory from machine learning, computational linguistics, and statistics.

Chapter 3 - Related Work We present related work to our project in this chapter. To avoid repeating the detailed accounts in the three papers, we will only account for any related work to our thesis en masse.

Chapter 4 - Results and Evaluation In this chapter we briefly account for the results and evaluations of the three select papers.

Chapter 5 - Conclusion This chapter concludes our thesis and present thoughts on further work.

Part II - Papers

Chapter 6 - Paper I This chapter presents the fist paper, *Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News*, in its entirety.

Chapter 7 - Paper II This chapter presents the second paper, *Optimizing Supervised Sentiment Lexicon Acquisition: Selecting Co-Occurring Terms to Annotate for Sentiment Analysis of Financial News*, in its fullness.

Chapter 8 - Paper III This final chapter presents the last paper, *The interaction impact of firm-specific news and market wide sentiment on stock price behavior - evidence from the Oslo Stock Exchange (OSE)*, in full.

Chapter 2

Theoretical Background

This chapter gives an overview of the central theories in artificial intelligence, hereunder machine learning (Section 2.1) and computational linguistics (Section 2.2), in addition to statistical methods (Section 2.3) comprising the foundation of this thesis.

2.1 Machine learning

In this section we account for machine learning theory which the sentiment engine construction in the thesis draws heavily on. First, we present five different feature selection processes in Section 2.1.1, then five different machine learning classifiers (Section 2.1.2), and, lastly, ways of evaluating such classifiers (Section 2.1.3).

2.1.1 Feature selection

Feature selection in text processing is the process of selecting a subset of attributes occurring in instances of the dataset and using only these as features in text classification. This process has two main purposes: 1) it simplifies training and classification by decreasing the input size, and 2) it increases classification accuracy by eliminating noise features[1] [Manning et al., 2008]. Accounts of five different feature selection techniques will now be presented.

[1] Features which, if included in the document representation, will increase the error when classifying new data. A model trained with noise features will be subject to overfitting.

Frequency-based

Frequency-based feature selection simply selects the attributes that are most common in the class. This can either be defined at the instance level (the number of times an attribute occurs in an instance) or at the dataset level (the number of instances that have attribute). This feature is computationally simple but has the fallacy that it can return attributes that are frequent and contain no specific information about a class [Manning et al., 2008].

Inverse-frequency-based

To mitigate this fallacy of frequency-based feature selection, the inverse-frequency-based method can be employed. This method acknowledges the fact that very frequently occurring attributes often are not very indicative of the contents of an instance and, hence, not well-suited for distinguishing between instances in a dataset. Examples of such attributes, if one considers simple terms, are stopwords, like *en*, *et*, *ett*, *det*, *dette*, *disse* etc, that will frequently occur in any instance without acting as a good proxy of its content [Papineni, 2001]. Furthermore, the inverse-frequency-based method de-emphasizes attributes that occur frequently in certain domains and, accordingly, have low distinguishing ability, like the terms *resultat* or *regnskap* in the financial news domain.

TF-IDF

A widely used algorithm for indexing in Information Retrieval is tf-idf, initially proposed by Jones [1972]. This selection method weights together the exhaustivity of an attribute, occurring frequently, with its specificity, occurring infrequently across the collection. Formally, the notation

$$tf\text{-}idf_a = tf_a \cdot lg\,(idf_a) \tag{2.1}$$

can be used to refer to the tf-idf statistic of attribute a. Here tf_a is the term frequency of the attribute and idf_a the inverse-document frequency, as previously described. TF-IDF can effectively be used to identify features that yield more value when a document is to be classified.

Mutual information

Mutual information is a measure of how much information an attribute contributes to making the correct classification decision on an instance. The mutual information of attribute t and instance c is calculated by:

$$I(t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(e_t, e_c) log_2 \frac{P(e_t, e_c)}{P(e_t)P(e_c)} \tag{2.2}$$

where the variable e_t takes the value 1 if the instance has attribute t and 0 if the instance does not have t. Correspondingly, the variable e_c takes the value 1 if the instance is in class c and 0 if the instance is not in class c. After computing $I(t,c)$ for all attributes t and instances c, one can select the k attributes with the largest values and, hence, arrive at the k attributes containing the highest mutual information. Denoting the set of the k attributes with the highest mutual as S, this formally becomes [Manning et al., 2008]:

$$S = \Big\{ t_1, t_2, ..., t_k \Big| \max_{t_1} I(t_1, c) \geq \max_{t_2} I(t_2, c) \geq ... \geq \max_{t_k} I(t_k, c) \geq \max_{t_{k+1}} I(t_{k+1}, c) \ \cap \ t_1 \neq t_2 \neq ... \neq t_k \neq t_{k+1} \Big\} \tag{2.3}$$

Chi-squared

In statistics, a χ^2 test is used to test for independence of two random variables. Two variables are independent if $P(A \cap B) = P(A) \ \ P(B)$, or equivalently $P(A|B) = P(A)$ and $P(B|A) = P(B)$. In feature selection the two random variables e_t and e_c annotate the occurrence of the attribute t and the occurrence of the class c, respectively (defined as in subsection 2.1.1). With the χ^2 feature selection method the attributes can then be ranked according to:

$$\chi^2(t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \tag{2.4}$$

where N is the observed frequency and E the *expected* frequency under the assumption that t and c are independent. In plain English, $\chi^2(t,c)$ measures the deviation of the expected counts E and observed counts N. A sufficiently high χ^2 value leads to the rejection of the null hypothesis stating that the random variables e_t and e_c are independent. This test statistic has only one degree of freedom and the probability of making an error, when performing the same test multiple times, becomes fairly high. Hence, one usually only assess the χ^2 statistics relatively. This means that they are only used to rank the (competing) attributes and select the k with the largest statistic. This is denoted by the set S like in subsection 2.1.1 [Manning et al., 2008]:

$$S = \Big\{ t_1, t_2, ..., t_k \Big| \max_{t_1} \chi^2(t_1, c) \geq \max_{t_2} \chi^2(t_2, c) \geq ... \geq \max_{t_k} \chi^2(t_k, c) \geq \max_{t_{k+1}} \chi^2(t_{k+1}, c) \ \cap \ t_1 \neq t_2 \neq ... \neq t_k \neq t_{k+1} \Big\} \tag{2.5}$$

2.1.2 Machine-learning classification algorithms

As previously discussed and illustrated in figure 1.3, machine-learning training and classification is the last step in turning an annotated dataset with extracted features into a sentiment analysis engine. Which machine-learning classification algorithm is selected, is paramount to the system's performance, both in terms of precision, running time and space requirements. Hence, a short introduction to machine-learning and accounts of some of the algorithms most relevant for our investigations are appropriate.

Machine-learning is the branch of Artificial Intelligence that is concerned with making computer programs automatically improve with experience, i.e. systematically using previously observed data to solve new encountered instance. A central problem in the fields of statistics and machine-learning is the problem of classifying examples into a discrete set of possible categories, known as the *statistical classification problem*. Formally, the problem task is to assign a target value v to each instance given this instance's attribute values $\langle a_1, a_2, ..., a_n \rangle$ [Anderson et al., 1986]. We will now account for five different machine-learning algorithms used to solve this problem in the domain of text processing, and hereunder sentiment classification.

Naïve Bayes

The Naïve Bayes classifier is a simple probabilistic model that applies Bayes' theorem[2] with the (strong) naïve assumption that the attribute values, $\langle a_1, a_2, ..., a_n \rangle$, are conditionally independent given the target value v. The Naïve Bayes classifier can formally be given by:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \tag{2.6}$$

where v_{NB} denotes the target value output by the classifier. To execute the algorithm a learning step is first performed where the various $P(v_j)$ and $P(a_i|v_j)$ terms are estimated, based on their frequencies in the training data, to arrive at a learned hypothesis. This hypothesis is then applied to new instances according to (2.6) [Anderson et al., 1986].

[2]Bayes' theorem gives the relationship between the probabilities of the two random variables A and B, the conditional probabilities of A given B, and vice versa: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Support Vector Machines

A Support Vector Machines (SVM) classifier is a computationally efficient way of learning 'good' separating hyperplanes in high dimensional feature spaces. This is done by finding the maximum margin separating hyperplane, given the training data mapped into this space, such that new points can be categorized with the highest degree of confidence, as illustrated in figure 2.1 [Cristianini and Shawe-Taylor, 2000].

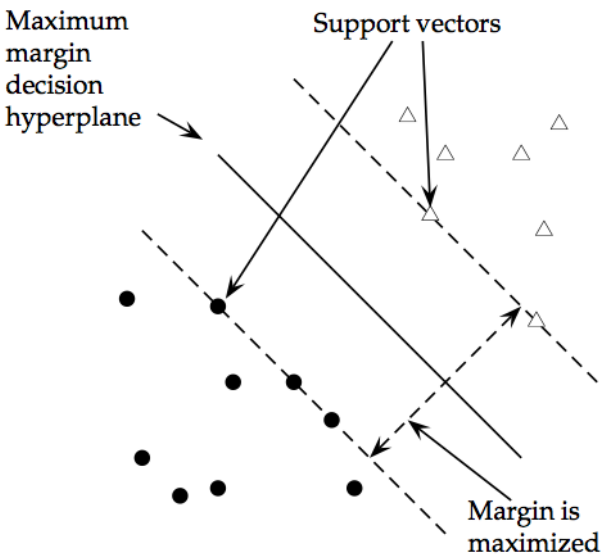


Figure 2.1: A maximum margin decision hyperplane with support vectors for classification of points in a 2-dimensional space (adopted from Manning et al. [2008])

It can be shown that given the instances $\{\vec{x}_i, y_i\}$ for $i \in P$ where $\vec{x}_i$ is a vector representation of the attributes, y_i is the classification and P is the set of points the problem of finding the support vectors for the maximum margin decision hyperplane can be reduced to solving a *quadratic optimization*[3] problem. The primal and dual representation of this (generic) optimization problem is given in Figure 2.2.

Quadratic optimization problems are well studied with numerous proposed

[3]A quadratic optimization problem has a quadratic objective function subject to linear constraints.

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\vec{w}^T\vec{w} && \text{maximize} \quad \sum\alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j \\ &\text{s.t.} \quad y_i(\vec{w}^T\vec{x}_i + b) \geq 1, \quad i \in P && \text{s.t.} \quad \sum_i \alpha_i y_i = 0, \\ & && \qquad \alpha_i \geq 0, \quad i \in P \end{aligned}$$

Figure 2.2: Primal and dual representation of SVM quadratic optimization problem

solving algorithms, including that of Megiddo and Tamir [1993] and Li and Zhang [2006]. The problem has, for certain instances, been shown to be solvable in polynomial time [Kozlov et al., 1980]. In addition to the generic solution outlined, SVMs can be extended to perform soft margin classification[4], solve multi-class problem instances and handle nonlinear cases. The latter is done by mapping points into a higher dimensional space, known as the *kernel trick* [Manning et al., 2008].

Artificial Neural Networks

Artificial Neural Networks (ANNs) are systems inspired by the human brain. It consists of interconnected neurons computing output from input values and letting information propagate through a (neural) network of such nodes. Perceptrons, the basic units of these networks, are functions that take a real-valued vector as input, calculate the linear combination of these inputs and output a 1 if the result is greater then some threshold. A visual illustration of a perceptron is shown in Figure 2.3.

Artificial Neural Networks consists of an input layer, composed of input nodes, one or more hidden layers, and an output layer, made up of output nodes. The hidden layers are essentially multiple perceptrons with edges between them. A sample ANN is shown in Figure 2.4.

Artificial Neural Networks can be constructed and trained by using the back-propagation algorithm which relies on the delta rule. This rule allows errors to propagate backwards in the network iteratively, adjusting the parameters of the network to arrive at satisfactory levels of precision [Mitchell, 1997].

[4]Soft margin classification adds slack variables to the optimization problem in order to trade off the overall target value with the misclassification of certain points.

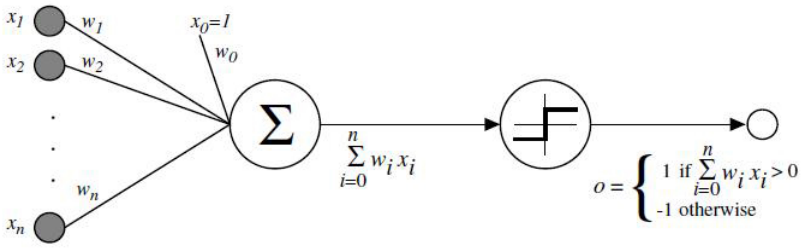


Figure 2.3: A perceptron adopted from Mitchell [1997]

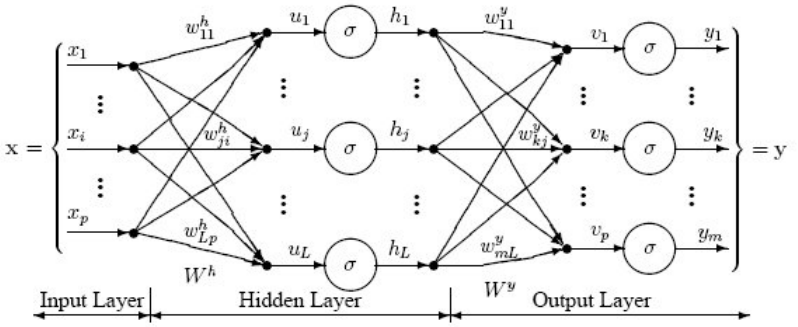


Figure 2.4: A sample ANN adopted from Furey [2012]

Random forest

The Random forest classification algorithm is an ensemble learning method first introduced by Breiman [2001]. This learning method, inspired by Ho [1995], combines the classification of several decision trees that are created during training, and outputs the average of the classifications made by each individual tree. Formally, the can be expressed as

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x') \tag{2.7}$$

where B denotes the number of trees used and $\hat{f}_b(x')$ is decision tree b's classification of instance x'. As the number of trees included increase the variance of the model decrease, thus resulting in a converged training and test error.

J48

J48, which we apply to achieve the highest classification precision in the financial Internet news domain considered in this thesis, is a Java[5] implementation of the more general C4.5 algorithm. This algorithm has the ability to generate decision trees for classification tasks. The algorithm was developed by Quinlan [1993] and built on prior work of the algorithm known as ID3 [Quinlan, 1986]. C4.5 builds decision trees based on a training set $S = \{s_1, s_2, ..., s_n\}$ where each element s_i in the training set has a vector $(x_{1,i}, x_{2,i}, ..., x_{p,i})$ of attributes. C4.5 construct trees in a recursive manner where nodes in the generated trees are created based on the attribute that yields the highest normalized information gain.

$$IG(A, S) = H(S) - \sum_{t \in T} p(t)H(t) \tag{2.8}$$

Information gain, as formulated in (2.8), measures the difference in entropy before and after C4.5 splits on an attribute. Here $IG(A, S)$ denotes the difference in entropy when the set S is split on the attribute A. $H(S)$ is the entropy of set S, T is the resulting subsets of the split on S, $p(t)$ denotes the proportion of elements in t compared to the elements in set S whilst $H(t)$ denotes the entropy of a subset t.

2.1.3 Classification evaluation

Once a classifier has been trained, an approach for evaluating the performance of it on a set of unseen records is needed. This is done to 1) determine the absolute error in classification and 2) to compare different classifiers working on the same dataset.

Holdout Method

The holdout method divides the original data into two disjoint sets. One set is used to train the classifier while the other is retained for testing. The training set and test set are usually a result of a 50-50 split on the original data. In some cases, depending on the size of the available data, only a third is reserved for testing. The classifier is evaluated by how well it performs classification of the retained test set. The major drawback of the holdout method is that it does not take full advantage of the dataset in training. Additionally, the model will also be dependent on the decomposition into training and test sets, since they are not independent being derived from the same dataset.

[5] A popular programming language, which the constructed sentiment engine is written in, created by Sun Microsystems.

Random subsampling

Random subsampling is a result of repeating the holdout method several times in attempt to improve the classifier. This iterative approach can be described with the formula:

$$acc = \frac{1}{k} \sum_{i=1}^{k} acc_i \tag{2.9}$$

where acc_i is the accuracy of the classifier in the i^{th} iteration and k is the total number of iterations. Random subsampling suffers from the same problem as the holdout method, since the entire dataset is never used in any training iteration. Another problem is that there is no way of knowing how many times a given instance has been used, either in training or testing, which could result in overfitting.

Cross-validation

Cross-validation, in comparison to random subsampling, makes sure that each record is used the same number of times for training, and exactly once for testing. This is done in a series of rounds where the dataset is partitioned differently into a training and test set each time. The average precision of the rounds is used to determine the overall accuracy of the classifier. Cross-validation, in comparison to the other introduced evaluation models, is better able to reduce the risk of overfitting. This is especially true when the training set is small and / or the classification model has a lot of parameters.

2.2 Computational Linguistics

Computational Linguistics is a discipline at the intersection of linguistics and computer science concerned with the computational aspects of natural language. Furthermore, linguistics is the scientific study of language; the interplay of sound and meaning [Halliday, 2006]. The latter, which is most interesting in the context of sentiment analysis, is concerned with how languages apply logic and real-world references to convey, process, and assign meaning, in addition to managing and resolving ambiguity. This, in turn, includes the study of semantics (conceptual inference from words) and pragmatics (contextual inference from meaning) [Chierchia and MacConnell-Ginet, 2000]. We will now define and account for some of the fundamental concepts in (computational) linguistics which are relevant for the construction of our sentiment engine.

2.2.1 Lexicon

A lexicon is a language's inventory of lexemes, the basic units of meaning roughly corresponding to words [Crystal, 2004]. In the context of sentiment analysis one needs lexica to maintain and assign mappings between lexemes in texts and sentiment values. With these mappings between lexemes, sentiment values and some way of aggregating the sentiment values of different lexemes holistic sentiment classifications of texts can be achieved.

2.2.2 Part of speech

Part Of Speech (POS), also known as word class, is a linguistic category of words [Kroeger, 2005]. Whereas the English language traditionally has been classified into eight word classes, the Norwegian language is divided into ten parts of speech [Bjørneset, 1999]. These classes are:

Substantiv (Noun) Names of places, persons and things

Adjektiv (Adjective) Description of one or more nouns

Pronomen (Pronoun) Replacement of noun

Determinativ Closer description of noun (has no corresponding word class in English)

Verb (Verb) Name of an action

Adverb (Adverb) Describes or modifies verb, adjective or other adverbs

Preposisjon (Preposition) Describes where the verb or noun is in relation to another verb or noun in both time and space

Konjunksjon (Conjunction) Combines two words from the same word class or two phrases

Subjunksjon Initiates phrases (has no corresponding word class in English)

Interjeksjon (Interjection) Special words expressing an emotion or sentiment on the part of the speaker

A POS-tagger is a software tool that automatically tags text into different parts of speech, also known as a tagset. These tend to operate with a tagset much larger than the wordclasses described in this section. For our further investigations we will rely on the Oslo-Bergen Tagger (OBT)[6]; a POS-tagger cooperatively

[6] A description and download of this POS tagger is available at www.tekstlab.uio.no/obt-ny/english/index.html.

developed at University of Bergen (UiB) and University of Oslo (UiO). It has a tagset of length 148 but for this investigation only the described classes are of interest to us. Hence, we will not go into the finer details of the OBT tagset here[7].

2.2.3 Pre-processing

Linguistic pre-processing is the preparation of a machine-processable representation of a word from particular instances of its occurrences [Pekar, 2004]. There are several pre-processing techniques that can be employed in computational linguistics and we will now account for the two known as lemmatization and tokenization in the two subsequent sections.

Lemmatization

Lemmatization, closely related to stemming[8], is the process of finding the normalized form of a word [Plisson et al., 2004]. An example of such a normalization is the reduction of the words *walked*, *walks*, *walking* to the base form, *walk*, which is called the lemma of the words.

Tokenization

Tokenization, which is a constituent of lexical analysis, is the process of forming tokens, strings of one or more characters that are significant as a group, from an input stream of characters [Webster and Kit, 1992]. Examples of tokenizations include the handling of digits, hyphens, punctuation marks, and case of letters.

2.2.4 Valence shifters

Valence shifters are (sets of) words that interact with other parts of a phrase or sentence altering its sentiment, polarity and/or strength [Polanyi and Zaenen, 2006]. These can be divided into five different categories.

Negatives Negatives are words that negate the polarity of a phrase or sentence. Examples of words in this category include *nei*, *ikke*, *ei*, *aldri*, *intet*. The sentences *Han er intelligent* and Han *er ikke intelligent* clearly have opposite sentiments.

[7] Full tagset with descriptions is available at www.tekstlab.uio.no/obt-ny/english/tagset.html.

[8] The difference between stemming and lemmatization is that stemming is algorithmically driven, while lemmatization is lexicon/dictionary-driven.

Intensifiers Intensifiers are words that increase or intensify the sentiment in a phrase, like *veldig*, *heller*, *riktig*, *mer*.

Diminishers Diminishers are words that decrease the sentiment in a phrase, e.g *lite*, *mindre*, *noget*, *noe*. The sentences *Han er veldig intelligent* and *Han er lite intelligent* have a stronger and weaker sentiment than the sentences without the modifiers *veldig* and *lite*, respectively. Interestingly, there are also interactions between negatives and intensifiers or diminishers that ideally would need to be accounted for. For instance, *Han er ikke veldig intelligent* has a weaker sentiment than *Han er ikke intelligent*.

Connectors Connectors are words that link phrases to form sentences. Often these phrases contrast or contradict each other. This category is composed of the conjunctions *og*, *for*, *men*, *eller*, *at*, *så* in addition to (multi-word) constructs like *samt*, *dessuten*, *i kontrast til*, *til tross for*, *på tross av*. For example, in the two-phrase sentence *Han er intelligent, men en drittsekk* the sentiment of the second phrase offsets that of the first equating to a negative overall.

Verbs Verbs can also be valence shifters since they are believed to have the strongest impact on overall sentiment, especially in short sentences, like titles or headlines. Verbs can also act as intensifiers or diminishers, like *økte*, *understreke*, *støtte* etc. In the sentence *Den intelligente mannen er mislikt* the latter verb controls the entire sentiment.

If one devices lexica for handling the different types of valence shifters introduced and constructs rules for handling these, one can (correctly) classify sentiment of texts with higher degrees of complexity [Simančík and Lee, 2009].

2.2.5 Co-Occurring Terms

Co-Occuring Terms (COTs) are terms occurring in the same context. They have the ability to capture semantic, lexical, or other relations between terms [Matsuo and Ishizuka, 2004] and therefore are suited as a base of sentiment lexica, as we do in Paper II of this thesis (Chapter 7). Formally, COTs can be defined as follows.

Definition 1 (Co-Occurring Terms)**.** *Co-Occurring Terms (abbreviated COTs) are terms that co-occur in the same text without being separated by any of the punctuation: periods, question marks, or exclamation marks.*

Consider, for instance, an article title that reads *Eier aksjer verdt 180 millioner. Kona får ny sykkel* (*Holds stocks worth 180 millions. Wife gets new bike*). From Definition 1 it follows that the terms *aksjer* (*stocks*) and *kona* (*wife*) are not COTs

being separated by '.', whereas the terms *aksjer* (*stocks*) and *verdt* (*worth*) are COTs. COTs are closely related to n-grams[9] but more generic in that contiguous appearances of words in text are not required. In the same sentence *aksjer* (*stocks*) and *millioner* (*millions*) are COTs but not an n-gram since the terms do not appear contiguously.

It is usual to add additional constraints when extracting COTs from documents, reducing the vast candidate space to consider. This is done by 1) confining the arity of COTs and 2) limiting the COTs radius.

Definition 2 (Arity)**.** *The arity of Co-Occurring Terms (COTs) is the number of terms they are composed of.*

In the above sentence the COTs *Kona får sykkel* (*Wife gets bike*) have arity 3 whereas *Kona sykkel* (*Wife bike*) have arity 2.

Definition 3 (Radius)**.** *The radius of Co-Occurring Terms (COTs) is the maximum allowed distance between the terms that are the furthest apart. The distance is the number of words between and including these outermost terms.*

In the same sentence, the radius of the COTs *Eier aksjer* (*Holds stocks*) is 2, the radius of the COTs *Eier verdt* (*Holds worth*) is 3, and *Eier millioner* (*Holds millions*) is 5.

2.3 Statistical methods

Statistical methods are procedures for analyzing or representing statistical data in addition to calculating statistics thereof [Miller, 1995]. In the first paper of our thesis we draw on statistical methods for assessing inter-rater reliability. These methods are detailed in Section 2.3.1. In our third and final paper we use the statistical methods including Autoregressive conditional heteroskedasticity (ARCH) regression models and the Wald hypothesis test, which are accounted for in Section 2.3.2 and 2.3.3, respectively.

2.3.1 Inter-rater reliability

Determining sentiments in the news domain is a subjective task. Therefore, ways of determining the reliability of the manually annotated dataset is needed. This is done by measuring the degree of consensus, or inter-rater reliability, between the different annotators who have annotated the same dataset. This can be done in several different ways, two of which will be accounted for in this section. Both of these were used to assess the annotation study we carried out as a part of Paper I (Chapter 6).

[9] An n-gram is a contiguous sequence of n terms in a text.

Joint-probability of agreement

The most basic approach for measuring inter-rater reliability is the joint-probability of agreement. This is simply the percentage of annotations where both (or all) annotators have agreed. This measure does not take the aspect of chance into account, i.e. if the two annotators come to an agreement only based on chance, this would not be reflected in the calculation. This could be a problem when there is a small number of annotation classes, resulting in a high likelihood of agreeing by chance [Uebersax, 1987]. Since we only use three annotation classes of subjectivity (articles are classified as either positive, neutral, or negative), this is definitely the case for our investigation.

Krippendorff's Alpha

Krippendorff's Alpha is derived from the generalization of several inter-rater reliability measures. It can be used for any number of annotators. It also takes into account incomplete data, all the values that can be used while annotating, as well as several levels of measurement. Additionally, the computed Alpha value can be compared across studies with different number of annotators, classifications, metrics and sample sizes [Krippendorff, 2012]. The Krippendorff's Alpha can, in its most generic form, be defined as:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2.10}$$

where D_o denotes the observed disagreement and D_e the expected agreement between the annotators, based on their use of the different annotation classes. These two measures can be calculated and estimated in various different ways according to the data and distributions in question. We will not account for these subtleties here and ask any interested reader to consult Krippendorff [2012].

2.3.2 ARCH model

The ARCH model was originally developed by Engle [1982], for which he later was awarded the Nobel price in Economics, and is used to characterize and model observed time series. This model assumes the variance of the current error term to be a function of the q previous time periods' error terms. The ARCH model has numerous extensions[10] but is in its original form defined as:

$$y_t | \mathscr{F}_{t-1} \sim \mathcal{N}\left(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}, \sigma_t^2\right) \tag{2.11}$$

[10] For an elaborate overview of ARCH model extensions, see Bollerslev [2008].

where $\mathscr{F}_{t-1}$ refers to the information set available at time $t-1$ and the conditional variance,

$$\sigma_t^2 = f\left(\epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-q}; \Theta\right) \tag{2.12}$$

is the explicit lagged function of the q lagged error terms, $\epsilon_t \equiv y_t - \boldsymbol{x^T\beta}$. This means that, for a model of order q, q lagged error terms with coefficients, $\alpha_i \epsilon_{t-i}^2,\ i \in \{1, 2, ..., q\}$, are added to the regression equation. These α coefficients, in addition to the β coefficients, are then estimated by the model through the maximization of a log-likelihood function of equation (2.11) for all t time periods [Bollerslev, 2008].

2.3.3 Wald test

The Wald test is a general parametric statistical test that can be used to test the true value of a parameter, θ, based on a sample estimate, $\hat{\theta}$. This is contingent that the data can be expressed as a statistical model with parameters to be estimated from the sample. When assuming that the difference between the estimate and true value is normally distributed and finding the maximum likelihood estimate of the standard error of the difference $\text{se}(\widehat{\theta})$, this difference can be compared to chi-squared distribution (with one degree of freedom),

$$\left(\frac{\widehat{\theta} - \theta}{\text{se}(\widehat{\theta} - \theta)}\right)^2 \sim \chi_1^2 \tag{2.13}$$

for the univariate case. The multivariate case, simultaneously testing multiple parameters, can be handled using a variance matrix [Wooldridge, 2012].

(This page is intentionally left blank.)

Chapter 3

Related work

In this chapter we will present some of the previous work related to our thesis. To avoid repeating the detailed accounts in the three papers, we will only describe any related work to our thesis en masse. The works mentioned in this section are, naturally, also related to the thesis's constituent papers. For elaborate, in-depth accounts of the related work of each paper, please see the relevant sections in these, which can be found in Chapter 6, 7, and 8, respectively.

There are numerous previously published works related to our thesis as a whole. At a high level these can be separated into works that 1) perform sentiment analysis in the financial domain (but do not get as far as linking it to stock price behavior), 2) perform sentiment analysis in the financial domain and simulate trading strategies to validate their approach (but without relying on / with less use of sophisticated statistical methods for validating their results), and 3) perform sentiment analysis and use well-justified regression models to reveal causal relationships between news and stock price behavior. These three categories of related work will now be accounted for in turn.

3.1 Sentiment Analysis of Financial News

In this first category, works include that by Devitt and Ahmad [2007] (sentiment analysis of news on the aggressive takeover bid of the low-cost airline Ryanair for the Irish flag-carrier airline Aer Lingus), Agić et al. [2010] (general finance news written in Croation), and O'Hare et al. [2009] (specifically for financial blogs).

3.2 Sentiment-Based, Simulated Trading Strategies

In this second category, previously published studies are Kim et al. [2014] (who reveal correlations between the sentiments published in Korean financial news and the Korea Composite Stock Price Index), Si et al. [2013] (who use financial tweets to predict short term movements in the S&P100 index), Zhang and Skiena [2010] (who document consistently favourable returns over the period 2005-2009 from a simulated trading strategy based on sentiment analysis of news and blogs relevant for the NYSE), Zhai et al. [2007] (find predictive power in joining news sentiments and aggregate-market wide sentiments in predicting movements of the mining and metals company BHP Billiton's stock, listed on the Australian Stock Exchange), Hollum et al. [2013] (who use a corpus of Thomson Reuter newswires collected from Dow Jones' Factiva for seven large stocks) in addition to Schumaker and Chen [2009] and Schumaker et al. [2012] (who use the developed text categorization engine AZFinText System[1] to predict stock price behavior 20 minutes after publication). We argue that these latter studies have weaknesses in that they 1) will be prone to overfitting (when trading strategies have been developed and tested on the same data) and 2) are much harder to evaluate for the world at large, when compared to the use of canonical statistical methods with transparent and completely reported results.

3.3 Statistical Methods for Using Sentiments to Predict Stock Price Behavior

In the latter and third category, we find the works of Ferguson et al. [2011], Uhl [2011], and Groß-Klußmann and Hautsch [2011]. The former uses sentiment analysis and regression analysis to link news flow to U.K. stock price behavior. Puzzlingly, the authors use Ordinary Least Squares (OLS) regression models, which, in the financial domain exhibiting time-varying volatility leads, arguably, to biased coefficient estimates. The two former works retrieve sentiment data from the NewsScope Sentiment Engine[2]. The classifications of this system are linked to stock price behavior using the robust (when compared to OLS) Vector Error Correction (VEC) models in Uhl [2011]. And in Groß-Klußmann and Hautsch [2011] the classifications are linked to high-frequency stock price data with the ability to trace the effect of news down to seconds after publication. Although both of these meticulous studies document the causal relationship between news

[1] This system is described at http://en.wikipedia.org/wiki/AZFinText.

[2] For (the scanty publicly available) information on this sentiment engine see www.thomsonreuters.com.

sentiment and stock price behavior, they do not 1) study the interaction impact of market-wide sentiment and firm-specific news on stock price behavior, as we do, and 2) have not developed the deployed sentiment engine in-house. This second part means that they are not easily able to extend their study to other stock exchanges where the news flow is written in a non-English language. Exactly this could be of great pecuniary interest, from a value investing perspective [Graham et al., 1934] and [Graham, 1959], since such stock exchanges, typically, are less efficient than those of the U.S. and the U.K. [Fama, 1970; Dickinson and Muragu, 1994]. Additionally, these studies rely on the commercially available sentiment engine offered by Reuters, and hence, one could argue that the competitive advantage of this system is limited since, in theory, anyone could deploy it without much effort.

(This page is intentionally left blank.)

Chapter 4

Results and Evaluation

This chapter presents the high-level research results of the thesis. Again, to avoid tedious reduplication of the results more elaborately displayed in the three papers, we will in this chapter only account for the most fundamental outputs of each paper. Each of these is a contribution to the research questions discussed in Section 1.3. These synopses are included to illustrate how the three papers constituting tge thesis tie together, comprising a coherent piece of work. This chapter is structured in the order that corresponds to research contributions discussed in Section 1.4. Specifically, we present the results of the sentiment engine construction in Section 4.1, before we describe the evaluation of this engine in Section 4.2. Lastly, we portray the results of studying interaction impact of firm-specific and aggregate market-wide sentiment on stock price behavior in Section 4.3.

4.1 Sentiment Engine Construction

In order to create a sentiment engine achieving satisfactory classification precision, we have in Paper I and Paper II addressed the two challenges outlined in the introduction (Section 1.2): 1) we have tailored sentiment analysis for the financial domain by investigating which feature categories are most important in classification, and, as a consequence of this, we have 2) solved the problem of the limiting availability of lexical and linguistic resources for the Norwegian language by optimizing a parsimonious method for manually constructing such a sentiment lexicon.

The sentiment analysis has been tailored for the financial domain by experimenting with different feature categories and evaluating how these contribute to classification precision. In doing so we manually annotated a corpus of 1000 news

articles from the financial online news publisher Hegnar.no[1], extracted features from these, in reliance on feature selection methods detailed in Section 2.1.1, and used the machine learning algorithms SVM, ANN, NB, RF, and J48, as delineated in Section 2.1.2, to build sentiment classifiers. These were then evaluated with averaged 5-fold cross-validation, as described in Section 2.1.3. The four different feature categories devised, and tested, were Textual (T), Categorical (C), Grammatical (G), and Contextual (X) and the ultimate classification precision results from running the five named classifiers on combinations of these categories are depicted in Figure 4.1. Note that the specifics of this work, in much greater detail than this high-level description, is given in paper I (Chapter 6).

Features Categories	**SVM**	**ANN**	**NB**	**RF**	**J48**
T†	57.4	57.7	**57.8**	56.7	57.3
TC†	57.1	57.8	**59.1**	57.6	57.9
TCG†	57.6	58.7	**59.0**	58.2	56.9
TCGX†	66.5	64.0	68.2	66.9	**68.7**
BEST‡	68.4	65.8	70.1	70.2	**70.8**

†Textual (T), Categorical (C), Grammatical (G), and Contextual (X) feature categories

‡Best feature composition per classifier

Table 4.1: Classification Precision Results by Feature Category and Machine Learning Classifier (%)

Table 4.1 clearly shows that Contextual (X) features yield classification performance enhancement when added to the other (less sophisticated) feature categories, and furthermore, that the classifier J48 seems superior in this domain. This finding formed the basis, as adumbrated in Section 1.5, for Paper II which addresses the importance of contextuality in sentiment analysis by studying methods for devising sentiment lexica with limited availability of lexical and linguistic resources, as is the case for the Norwegian language. This was done by the means of a parsimonious approach, relying on manual annotation, where we optimize a number of parameters in this lexicon construction. These parameters include 1) which radius (r) should be used when extracting Co-Occurring Terms (COTs) (as detailed in Section 2.2.5) from an article collection, 2) which ranking function (f) should be used to order the candidates for entry in the lexicon, 3) how many such candidates (ρ) should be manually annotated and stored in the lexcion, and, lastly, 4) which machine learning classifier (c) should be used in the final classification. Figure 4.2 presents the classification precision results from varying these four parameters ($r \times f \times \rho \times c$).

[1] www.hegnar.no

		$\rho = 10\%$				$\rho = 30\%$				$\rho = 50\%$			
MLC	**Ranking**	**COT radius =**				**COT radius =**				**COT radius =**			
	Function	**2**	**4**	**6**	**8**	**2**	**4**	**6**	**8**	**2**	**4**	**6**	**8**
NB	*tf*	63.9	55.7	55.5	58.0	63.6	65.4	65.3	65.7	64.8	68.3	66.6	65.2
	idf	62.8	62.8	62.8	62.8	63.8	62.8	65.1	64.8	65.9	64.2	65.6	67.4
	tfidf	63.0	57.8	56.3	58.7	62.6	65.3	65.1	66.7	64.3	68.1	66.7	67.3
	mi	59.3	61.6	59.1	60.5	61.8	66.0	65.6	65.4	66.0	68.5	68.9	68.9
	χ^2	62.8	60.0	60.7	62.8	59.6	61.4	60.8	60.4	59.4	57.5	60.9	65.4
RF	*tf*	64.7	62.4	62.9	63.7	64.0	65.5	66.6	65.6	65.7	69.0	66.6	67.0
	idf	62.8	62.8	62.8	62.8	63.8	62.8	65.0	65.3	66.2	64.9	66.4	66.9
	tfidf	64.4	63.2	62.9	63.6	64.2	66.4	65.0	66.7	65.4	68.9	66.4	67.0
	mi	62.8	63.3	64.5	65.2	62.7	65.6	66.4	65.4	65.2	68.7	68.9	68.7
	χ^2	62.8	62.8	62.8	62.8	62.8	62.1	62.6	63.1	62.7	64.6	63.9	65.1
J48	*tf*	65.1	62.7	62.7	63.7	63.8	65.9	66.3	65.6	65.9	68.9	66.8	67.6
	idf	62.8	62.8	62.8	62.8	63.8	62.7	64.9	65.3	65.5	64.8	66.6	67.4
	tfidf	64.4	62.9	62.9	63.3	63.9	66.5	64.7	66.9	66.2	69.0	66.5	67.6
	mi	62.7	63.4	64.7	65.2	62.7	65.6	65.9	65.9	65.4	68.7	**69.1**	68.9
	χ^2	62.8	62.8	62.8	62.8	62.8	62.8	62.6	63.0	62.6	64.8	63.9	65.0

Table 4.2: Classification Precision Results by Machine Learning Classifier (c), Ranking Function (f), COT Radius (r), and (Relative) Lexicon Size (ρ)

This figure shows two important things: 1) that state-of-the-art classification precision is possible to achieve through this parsimonious approach (the classifiers have not been tuned per parameter configuration out of run time consideration, and if done so, would have yielded even higher precisions) as the state-of-the-art classification precision for this highly subjective task is ∼70% [Balahur et al., 2010; Mourad and Darwish, 2013], and 2) the optimal parameter values. Specifically, a larger sentiment lexica diminishingly increases precision, the radius should be kept relatively small, that the ranking function mutual information and machine learning classifier J48 should be employed.

These findings, in extension of the finding of paper I, formed the bases, as outlined in Section 1.5, for Paper III. In this paper, we firstly seek to evaluate this sentiment engine developed using methods from the disciplines machine learning and computational linguistics. Note that the results given in this section are described in much more elaborate detail in paper II (Chapter 7).

4.2 Sentiment Engine Evaluation

With the sentiment engine for classification of financial Internet news articles, achieving state-of-the-art precision, the ultimate test to this system is to see whether this can make predictions on stock price behavior. Using ARCH regression models and simple Z hypothesis tests with Bonferroni correction[2] we are

[2]Bonferroni Correction states that rejecting all $p_i < \frac{\alpha}{n}$ will control that the familywise error rate will be below α [Abdi, 2007].

successfully able to link positive, neutral, and negative news article publication count, as classified by our sentiment engine, to the three dependent variables 1) return, 2) volume (traded monetary value), and 3) order size. This was done for ten different stocks[3] listed on the OSE. These relationships are illustrated in Figure 4.1. These subfigures show simple regression models estimating the effect of positive, neutral, and negative article publication count to return, volume and order size. The same charts also show sample mean and standard deviation value by article count in the form of whiskers, hoping to illustrate the general relationships between these variables. These simple regression models have aggregated all the ten stocks together, which from a statistical standpoint is an imprecise simplification and, hence, unfit for exact statistical inference. Nevertheless, we argue this depiction still has communative value in illustrating the general relationships of the variables in question. Paper III (Chapter 6) employs more sophisticated statistical methods to make these inferences, but will not be accounted for in this section due to their convolution and prolixity, unfit for this high-level synopsis.

4.3 Interaction impact of firm-specific and aggregate market-wide sentiments on stock price behavior

With sentiments being the topic of this thesis and a classification engine well-evaluated and in place, we took our analysis further by examining firm-specific and aggregate market-wide sentiments to see if the interaction between these two variables had any predictive value on stock price behavior. In order to measure aggregate market-wide sentiments we used Principal Component Analysis [Jolliffe, 1986] to aggregate seven sentiment proxies[4] into a single index. This index, as an eyeball test, is graphed with the Oslo Stock Exchange All Shares Index (OSEAX) in Figure 4.2. From the figure it seems like the developed index correlates fairly well with OSEAX; it plunges upon the outbreak of the financial crisis late 2008 and resurges slowly thereafter, in accordance with the OSEAX.

[3]The tickers of these ten stocks are FUNCOM, IOX, NAUR, NOR, NSG, RCL, SDRL, STL, TEL, YAR.

[4]The sentiment proxies aggregated were 1) number of IPOs in the period, 2) average IPO return on first day of issue, 3) retail fund flow ratio (capital inflow divided by capital outflow for the period), 4) insider trade filing count, 5) total volume (in monetary value) on OSE, 6) ratio of newly issued equity relative to bonds by stocks listed on the OSE, and 7) price-to-book value for OSE combined.

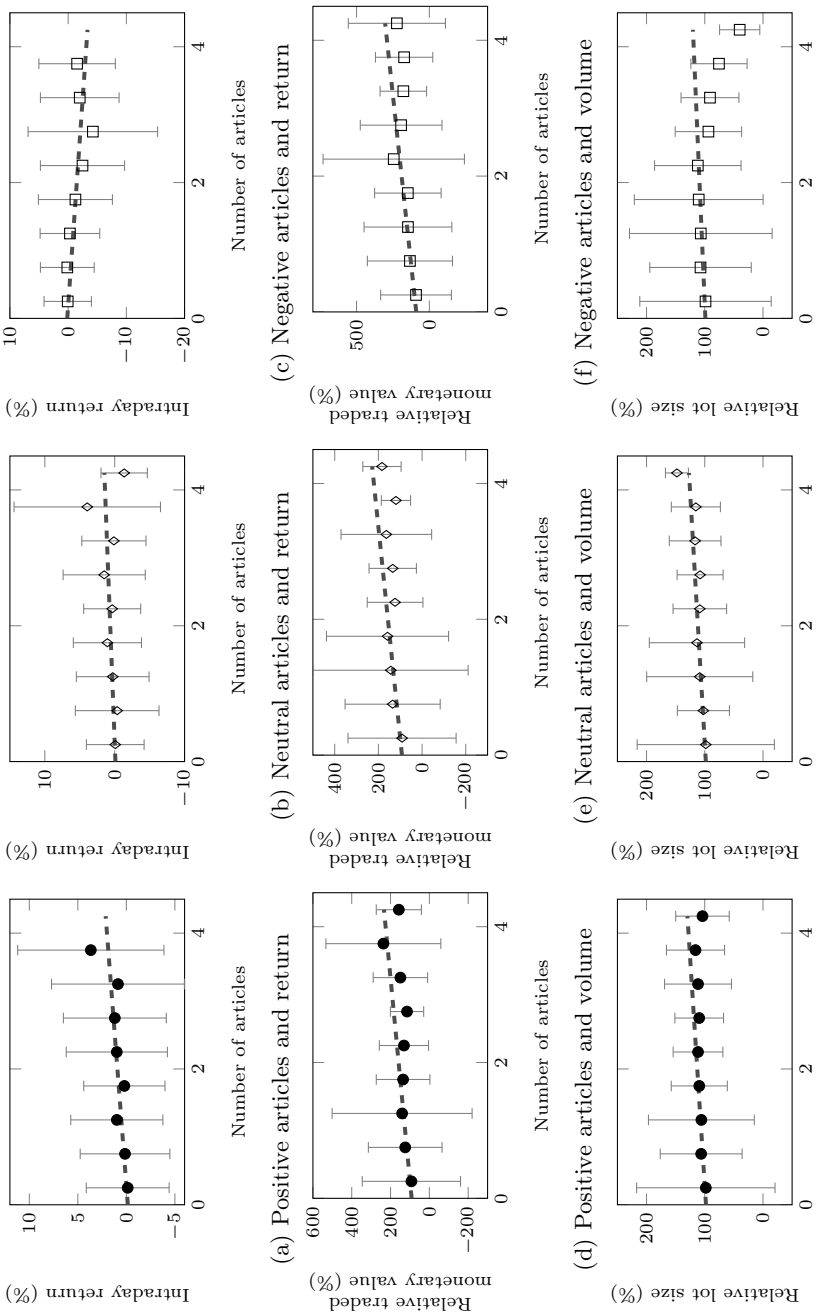


Figure 4.1: Relationships between the number of published articles, by classification, and intraday return (a, b, and c), relative traded monetary value (as a percentage of daily average) (d, e, and f) and order size (as a percentage of daily average) (g, h, and i)

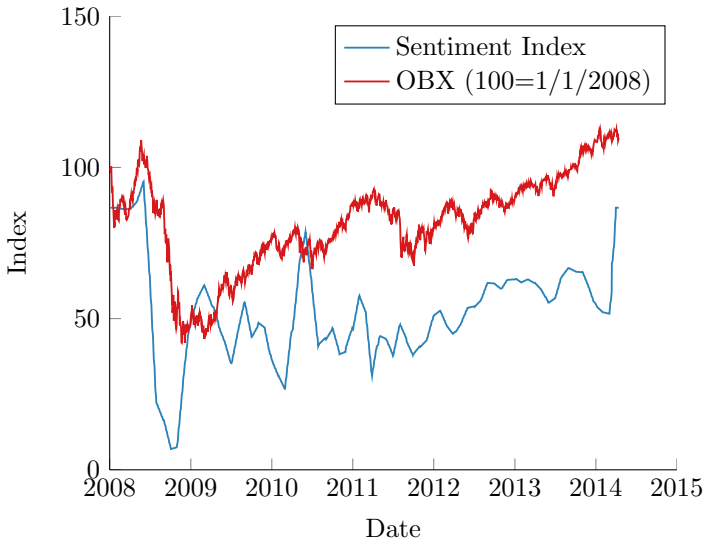


Figure 4.2: Historical development of Sentiment index and OSEAX. Both have been rebased to 100 at the 4^{th} of January 2008 (first day of trading in 2008).

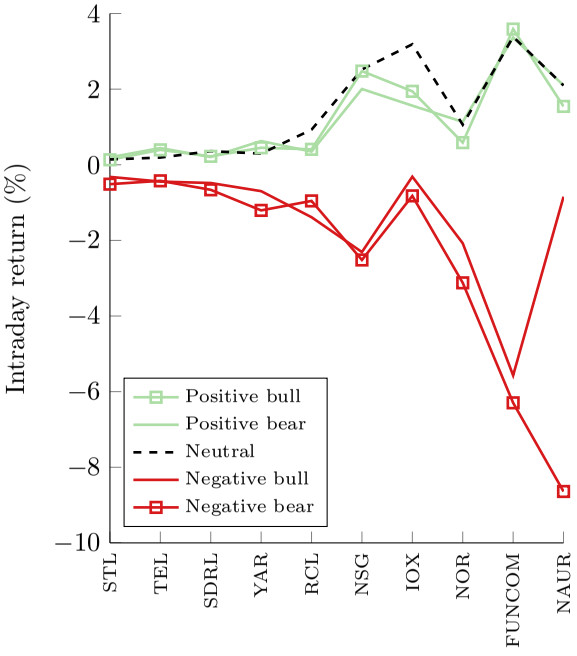
Using ARCH regression models and several different hypothesis tests we revealed statistically significant relationships between the interaction of firm-specific news, as classified by our devised sentiment engine, and this aggregate market-wide sentiment index on the three variables 1) return, 2) volume (traded monetary value), and 3) order size. In Figure 4.3a and 4.3b we have graphed the estimated β coefficients (output from the ARCH models) for the ten different stocks considered for positive, neutral, and negative firm-specific news, in interaction with aggregate market-wide sentiment[5], for intraday return and average weekly order size, respectively. These figures show that the effect of published firm-specific news sentiments on return and order size is different for 'Bull' and 'Bear', although the general sign of this effect remains somewhat unclear. From the figures, however, it seems that white chip stocks (the five rightmost stocks on the x-axis, which has been sorted by liquidity) react more strongly to news than blue chip stocks (the five leftmost stocks).

In Figure 4.4a and 4.4b we have again graphed the estimated β coefficients for the ten different stocks considered for positive, neutral, and negative firm-specific news, in interaction with the same aggregate market-wide sentiment,

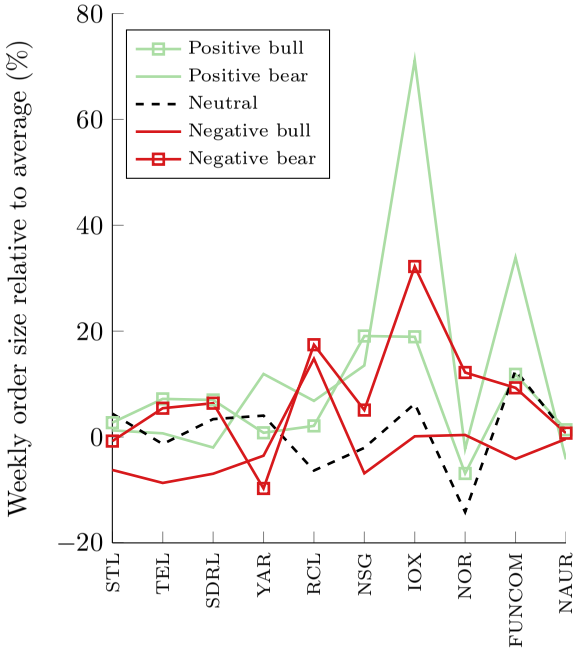
[5]'Bull' indicates a positive aggregate market-wide sentiment (i.e. the index is above average) and, conversely, 'Bear' indicates a negative aggregate market-wide sentiment (index below average).

this time for weekly and monthly traded monetary value, respectively. These figures show, like for daily return and average weekly order size, that the effect of published firm-specific news sentiments on volume is different for 'Bull' and 'Bear' sentiments, although the general sign of this effect remains somewhat unclear. These figures also seem to indicate that white chip stock reactions to news have greater magnitude than that of blue chip stocks.

Note again that this section is merely a synopsis of the investigation on interaction impact of firm-specific and aggregate market-wide sentiment on stock price behavior, and that the intricate details of this is accounted for in full in paper III (Chapter 8). In coda, these novel findings are, if not surprising, valuable for any system attempting to automatically condition stock trades on firm-specific sentiments; factoring aggregate market-wide sentiments into such a trading model seems justified and worthwhile.

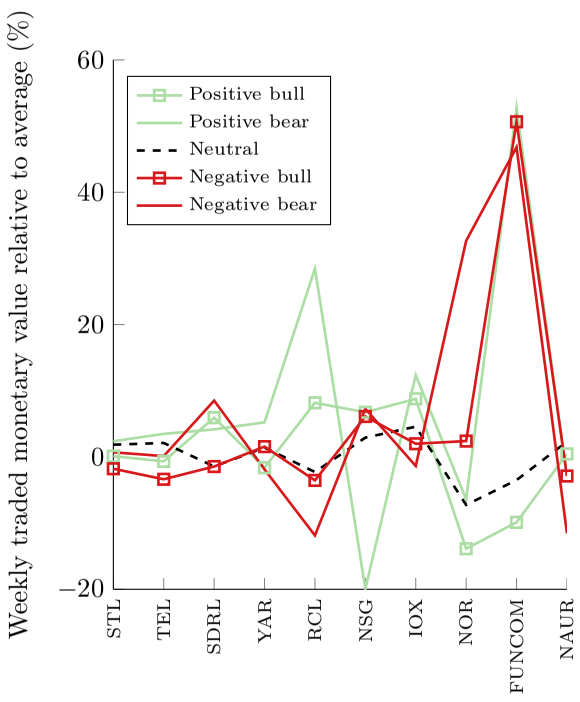


(a) Estimated β coefficients for predicting daily return in interaction with sentiment index

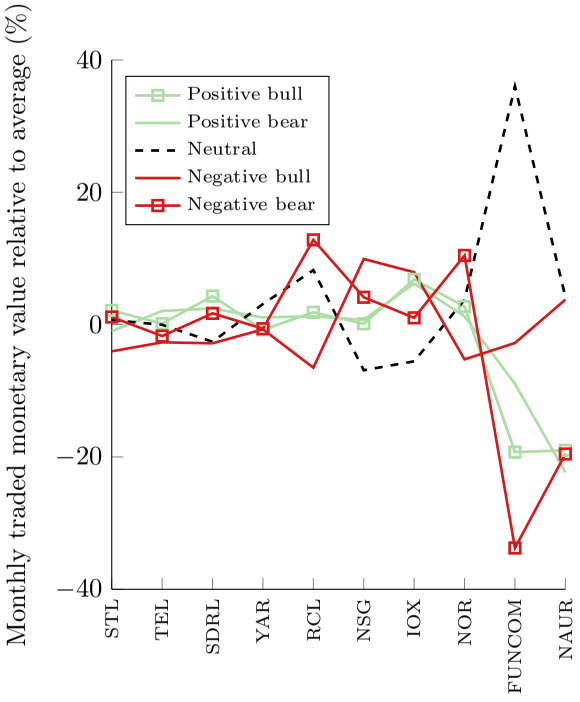


(b) Estimated β coefficients for predicting weekly order size in interaction with sentiment index

Figure 4.3: β coefficient graphs predicting return and order size



(a) Estimated β coefficients for predicting weekly traded monetary value in interaction with sentiment index



(b) Estimated β coefficients for predicting monthly traded monetary value in interaction with sentiment index

Figure 4.4: β coefficient graphs predicting traded monetary value

(This page is intentionally left blank.)

Chapter 5

Conclusions

In this chapter we present conclusions of the research conducted within the scope of this thesis by summarizing its contributions (Section 5.1) and discussing compelling venues for further work (Section 5.2).

5.1 Summary of Contributions

The overall topic of this thesis, as deliberately designated by the title, is to use sentiment analysis, a classification task within the field of artificial intelligence, for financial applications. Hereunder, we combine machine learning, computational linguistics, and statistical methods for the application of predicting stock price behavior of shares listed on the Oslo Stock Exchange (OSE). This was done based on the publication of firm-specific news articles and a devised sentiment index. The motivation for this approach comes from, as discussed in Section 1.1, news being viewed as a most felicitous source of information; effectively acting as a filtering, aggregating, and widely-read funnel of sentiments. Furthermore, the OSE was selected as the financial market of investigation, firstly, for its faculty of being inefficient, when compared to peer marketplaces, and, secondly, for the inherent linguistic barriers to processing the natural language associated with the exchange, being written in the linguistically resource-meager Norwegian language. If able to surmount these barriers, one could potentially attain a competitive advantage when trading in this market. In response to the research questions, listed and detailed in Section 1.3, this thesis has four main contributions, which now will be accounted for in turn.

In answer to the first research question of this thesis, we have found that contextual feature categories are paramount in performing sentiment classification of financial Internet news articles written in the Norwegian language. This was

the main finding of the first paper that is presented in this thesis. The primacy of contextual features in sentiment classification of financial Internet news articles necessarily instituted the need for finding methods for devising contextual sentiment lexica. This naturally comprised the foundation for our second research question, addressed in the second paper of this thesis.

In response to this latter, and second, research question, we found that sentiment lexica for the Norwegian language, with its named meagre lexical and linguistic resources, can be devised through manual annotation at permissible levels of efforts. All the while, this parsimonious approach achieves state-of-the-art classification precision. In optimizing the parameters input to this lexica construction, we found that COT radius employed should be kept low, that the ranking function mutual information should be used, that the lexicon size should be at least ∼30% of the COT candidate list, and that the machine learning classifier J48 should be applied in optimum. The optimized parameters output from answering this research question were finally used as input to the final sentiment engine used to classify the news flow associated with ten selected stocks listed on the OSE.

This sentiment engine classifying financial Internet news articles written in Norwegian with state-of-the-art performance, in reply to the third research question, was found able to make statistically significant predictions on the stock price behaviors return, volume, and order size. Positive articles were found to, predominantly, lead to significant increases in volume while negative articles were predicted to have the opposite effect. The same was found to be the general proclivity for order size. For return, only negative articles were found to significantly impact future stock price behavior and the publication of such articles were largely found to, ceteris paribus, reduce subsequent returns.

The interaction between news articles and market mood was also revealed to be statistically significant, in answer to the fourth and final research question. Although the sign of this effect seems firm-idiosyncratic, our analysis revealed that white chips' reactions are of greater magnitude than that of blue chips.

This concluding synopsis, in avoiding repetition of the accounts in the three papers less the very main contributions, has left out several noteworthy findings detailed in these papers, which are presented in full in Chapter 6, 7, and 8.

5.2 Further Work

We have identified, will now account for, and discuss four interesting venues for further work. Each of these is an extension of one of the research questions, and accordingly a contribution, of the thesis:

- In reply to the first research question, we now know that contextual fea-

tures are paramount in sentiment classification of financial Internet news articles written in the Norwegian language. Two interesting venues for further work, in extension of this finding, is to examine whether 1) drawing on more sophisticated linguistic theory (like diving deeper into the semantic interplay of words and sentences) in feature extraction and 2) developing more granular sentiment classifiers (we could, for instance, categorize texts down to industry or company level and train separate classifiers for each category) lead to sufficient performance improvement justifying the increased effort.

- Much research in sentiment lexicon acquisition is concerned with extending preexisting lexica with new entries and automatically classifying these [Feldman, 2013]. In response to the second research question of this thesis, we have devised, and optimized, a parsimonious method for constructing sentiment lexica. This was done for the Norwegian language, alas its meagre lexical and linguistic resources, by the means of manual annotation. It would now be interesting to investigate whether coupling some of the preexisting algorithms on sentiment lexica extension with our developed methods would lead to either 1) enhanced classification precision performance or 2) reduced required manual annotation effort.

- We have, in answer to the third research question of this thesis, validated our sentiment engine's ability to predict stock price behavior. A natural next step from this is to enhance the system to also output trading recommendations. The transactions produced by the engine can then be evaluated by simulating the returns of these over time. There are, however, a number of details that need to be taken care of before this can be done in practice, including 1) when trades should be recommended in timely relation to news publication (which we need higher resolution data in order to handle appropriately), 2) with which volume, and 3) at which quoted prices these recommendation should be made (also necessitating data on on a share's order book over time). In coda, even though the validation of our sentiment engine is promising, there are series of hurdles that need to be overcome before real-time deployment is possible.

- To be able to draw stronger conclusions on the interaction impact of firm-specific news and aggregate market-wide sentiment on return, which was the topic of the fourth research question, we plan to extend our initial analysis with high-frequency data. Additionally, broadening our investigation to include more stocks, if not with all, of the stocks listed on the OSE could lead to stronger, more general conclusions. And lastly, comparing this relationship between news sentiments and stock price behavior across stock

exchanges could also be of great interest as this could shed some light on which exchanges should be the target of news sentiment trading strategies.

Part II

Papers

(This page is intentionally left blank.)

Chapter 6

Paper I

Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, Wei Wei, and Jon Atle Gulla: *Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News*, to appear in the Proceedings of the 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2014).

(This page is intentionally left blank.)

Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News

Pål-Christian S. Njølstad*, Lars S. Høysæter*, Wei Wei† and Jon Atle Gulla*
*Department of Computer and Information Science
Norwegian University of Science and Technology
†Microsoft Development Center Norway
{palchrnj, larssmor}@stud.ntnu.no, weiweikth@gmail.com, jag@idi.ntnu.no

***Abstract*—Work on sentiment analysis has thus far been limited in the news article domain. This has mainly been caused by 1) news articles lacking a clearly defined target, 2) the difficulty in separating good and bad news from positive and negative sentiment, and 3) the seeming necessity of, and complexity in, relying on domain-specific interpretations and background knowledge. In this paper we propose, define, experiment with, and evaluate, four different feature categories, composed of 26 article features, for sentiment analysis. Using five different machine learning methods, we train sentiment classifiers of Norwegian financial internet news articles, and achieve classification precisions up to ∼71%. This is comparable to the state-of-the-art in other domains and close to the human baseline.**

Our experimentation with different feature subsets shows that the category relying on domain-specific sentiment lexica ('contextual' category), able to grasp the jargon and lingo used in Norwegian financial news, is of cardinal importance in classification - these features yield a precision increase of ∼21% when added to the other feature categories. When comparing different machine learning classifiers, we find J48 classification trees to yield the highest performance, closely followed by Random Forests (RF), in line with recent studies, and in opposition to the antedated conception that Support Vector Machines (SVM) is superior in this domain.

***Keywords*-Text analysis, Web mining, Feature extraction, Machine learning, Supervised learning, Support vector machines, Artificial neural networks, Decision trees**

I. Introduction

With the proliferation of internet access and data accessibility, the need for tools to effectively search all the available information has become paramount. The magnitude of this need is evident in the search engine corporation Google's second-to-none part of peoples' internet lives and its market capitalization; making it the world's second largest company (as of February 2014). Traditional search technologies, like Google's, however, usually only reflect explicitly contained document information. Analysis seeking to exploit implicit information is much less researched, like the sentiments expressed in an internet news article. Traditional keyword based search strategies are, for instance, not able to reveal the presence, nor polarity, of such expressions.

Most sentiment analysis, that precisely tries to identify such implicitly contained information, has traditionally been conducted within the intrinsically sentiment-bearing domain of product reviews (in addition political speeches and blogs). English has remained the predominant language of study. Research in this domain continues to exhibit improvements. Recently, [9] claimed to push the state-of-the-art single-sentence product review sentiment classification precision from ∼80% to ∼85%.

The same cannot be said for sentiment analysis of news articles. For the aforementioned reasons, this domain endures as a much harder nut to crack. However, certain sub-domains appear more suited for sentiment analysis than others, such as financial news. This subdomain is 1) prone to bear sentiments (e.g. subjective predictions about the future are commonplace), 2) has clearly defined targets that these sentiments are expressed towards (like financial tradable entities) and, 3) is written, in ease more than necessity, with little variations in features, form and style of language. These three factors, coupled with the potential remunerative rewards, if able to automatically link financial tradable entities' price developments and news sentiments, make sentiment analysis compelling within this domain. Furthermore, the Norwegian language is suited for our research purposes since it is the language of the news flow covering the stocks listed on the Oslo Stock Exchange (OSE). Compared to peer financial marketplaces, the OSE is both more volatile and less efficient, leaving greater opportunities for exploiting potential predicative relations between the sentiments expressed in a stock's news flow and price development [10]. Additionally, sentiment analysis in this language is valuable since it previously has not resulted in any published research. For these reasons, we have focused our investigation to the domain of financial news written in the Norwegian language. Our main contribution in this paper is the evaluation of different feature categories highlighting the cardinal importance of contextual features in the domain of Norwegian financial news. We propose and define four different feature categories (Textual, Categorical, Grammatical and Contextual), having increasing degrees of complexity, and experiment with and evaluate these using five different machine learning classifiers. We find that the feature category relying on context-specific sentiment lexica, able to grasp the jargon and lingo used in Norwegian financial news, to contribute the by far most to classification - yielding a precision increase of ∼21% when added to the other feature categories. Additionally, J48 classification trees were found to be the highest

performing classifier attaining precisions up to ∼71%, comparable to the state-of-the-art and close to the human baseline. This is in line with more recent research claiming decision trees to be superior in sentiment classification [5], [6], [7].

We will now outline the remainder of the paper. A brief account of some of the related work to our paper is given in section II. Followingly, the three main parts of our research are accounted for in the three subsequent sections:

- *Annotation study* - Since no pre-annotated dataset has been available for our purposes, we carried out an annotation study manually annotating articles from the online news publisher hegnar.no. The results of our annotations were assessed with inter-rater reliability metrics to ensure their adequacy for further analysis. This is described in section III.
- *Feature engineering* - As input to the machine learning methods to be used in training and classification, we extracted 26 features from these annotated articles. These features were grouped into four categories with different levels of complexity. The definitions, examples and further details of these feature categories are given in section IV.
- *Classification training and testing* - Lastly, we train, test and compare different machine learning classifiers with varying feature subsets. The classification results for different feature subsets are accounted for and discussed in section V.

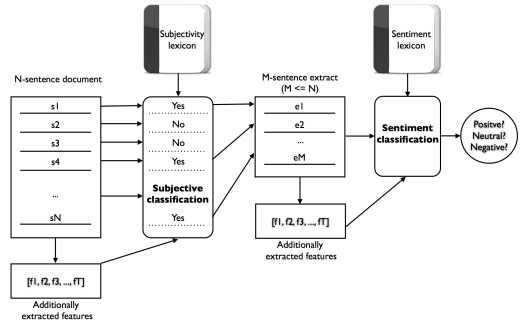
Finally, we conclude and briefly discuss interesting venues for future work, in section VI.

II. Related Work

Sentiment analysis is conducted either at the word, phrase, sentence, paragraph or document level, and one typically distinguishes between supervised or unsupervised approaches [11]. Since we're using machine-learning on an annotated dataset to attain aggregate article classifications, our focus in this related work section will be directed towards supervised approaches at the document-level. Additionally, since our main contribution is the evaluation of feature sets and machine learning classifiers, we'll also discuss some related work on feature selection for sentiment analysis using such classifiers.

In an attempt to summarize the related work to this paper, we have in Figure 1 depicted a model of a generic supervised approach to document-level sentiment analysis, influenced by [12]. Firstly, if considering a document at the sentence level, one can through subjectivity analysis classify each of the document's constituent sentences as either subjective ('Yes' in the Figure) or objective ('No'). In doing so, a subjectivity lexicon is usually employed. Additional features of the document may also be extracted to aid in this subjectivity classification task. Next, the sentences classified as subjective are input to a sentiment classification model for further analysis. This will, in turn, typically employ a sentiment lexicon to classify each sentence before these are aggregated to achieve a

Figure 1: Generic supervised document-level sentiment analysis approach using machine learning



document-level sentiment classification. Moreover, additionally extracted document features can serve as input to this classification engine. Since subjectivity analysis is, by in large, merely sentiment analysis less the polarity classification, many systems omit this intermediary step, as we do in our approach.

Related work differ in data set used, feature extraction, deployed lexica, and employed machine learning classifier, among others. Later research has leveraged yet more sophisticated natural language processing techniques, like valence shifters [13] and combinatory categorial grammar [14] in feature extraction. Numerous machine learning classifiers have been applied in sentiment analysis. Although Support Vector Machines (SVM) generally has been perceived as the superior classifier in text processing, and hence sentiment analysis, recent studies have found Random Forrest and J48 classification trees to perform the best in this particular classification task [5], [6]. We will now account for the specifics of some related work and their classification precisions, where appropriate.

In [12], a minimum-cut framework with a Naïve Bayes (NB) classifier was used to determine the polarity of review extracts. [4] used the discounted NEAR operator and PMI-IR algorithm to achieve an average classification precision of ∼74% in a similar domain. [3] experimented with different machine learning frameworks on movie reviews, achieving the highest classification precision of ∼83% with SVM. Artificial Neural Network (ANN) have also been applied to the same task, like in [15].

As previously noted, the news domain is both less researched and understood, evident in [2]. These authors experiment with classifying documents from different sources attaining precision results between ∼75% and ∼95%. Particularly, the developed framework struggled with news article documents, yielding precisions down to ∼75% due to difficulty in dealing with long and complex sentences, which are common in news articles.

Perhaps the most relevant previous work to ours is [1]. The authors discuss the definition of sentiment in the context of news articles and advocate a redefinition of this to include only explicitly stated sentiment expressions in the text, referred to as textual sentiment. With this

Table I: Inter-Reader Reliability Results (%)

Article part	Simple probability of agreement	Krippendorff's alpha
Title	75.4	69.7
Lead	73.5	63.5
Main	62.8	43.9
Aggregate	67.8	70.3
All parts*	43.2	73.2

*Agreement on title, lead, main and aggregate annotations

(re)definition, and an annotation guide, the authors are able to achieve a simple probability of agreement of 81% in annotation of 1592 quotes extracted from news articles. Furthermore, they experimented with different lexica in sentiment analysis and achieved classification precision up to 82%. In comparison to our research of classifying aggregate article sentiment, however, this must be considered a less arduous task since the analysis is only done at the single sentence-level.

Recently, numerous approaches to improving the feature selection process in sentiment analysis have been made. [16] presented a feature subsumption hierarchy, which both simplifies and removes unnecessary features. Using a similar approach, [17] proposed a rule-based multivariate text feature selection method called Feature Relation Network (FRN). Both of these studies focused on n-gram features. [18] address the same problem, but generalize their approach to substring-group features as oppose to n-grams. In [19] information gain and genetic algorithms were used as heuristics in feature selection - related to our approach of using Mutual Information, χ^2 and the Maximal information coefficient [20] in guiding which candidate features to perform exhaustive searches for optimal subsets over.

Interested readers are asked to consult [11] for a broader and more comprehensive overview of recent work in sentiment analysis.

III. Annotation Study

As the source of our dataset, we selected the online financial news publisher hegnar.no, and extracted 1000 articles from its website. This was in effect all news published in the time period from Feb 4^{th} to March 26^{th} 2013.

Since the definition of sentiment is inherently subjective, it is preeminent that there is a common understanding of what constitutes sentiment when annotating. Like [1], we relied on the purely textual sentiment definition, as this can be thought of representing a common denominator of all readers (annotation without the reliance on background knowledge and / or any preconceived biases). To aid the annotation task, and to facilitate further extensions of the dataset, we devised an annotation guide with detailed instructions, and example annotations. This is available upon the reader's behest and contact of the corresponding author of this paper.

The articles to be annotated were composed of three parts: title, lead and main text. Each of these were handled separately, in order of appearance. Lastly, an aggregate sentiment classification for the article was also given. No specific instructions on how to weight the different constituting parts together arriving at the aggregate classification were given in advance, remaining at the holistic discretion of the annotators. Each of the article parts, and the aggregate, were annotated with either the negative (*-1*), neutral / objective (*0*) or positive (*+1*) sentiment classification. Albeit some research, like [21], emphasizes that sentiment-neutral expressions need not be objective, like the word *surprise*, we were not trying to distinguish this special case.

Table II: Confusion Matrix 'Aggregate' Annotations (row annotations by annotator 1 and column by annotator 2)

	-1	0	1
-1	184	38	6
0	97	257	127
1	10	44	237

Table III: Correlation Matrix on Articles with Agreement on All Parts

	Title	Lead	Main	Aggregate
Title	1.000	0.748	0.599	0.886
Lead	0.748	1.000	0.659	0.826
Main	0.599	0.659	1.000	0.755
Aggregate	0.886	0.826	0.755	1.000

We evaluated the inter-rater reliability of our annotations with both the 'simple probability of agreement', and the more robust 'Krippendorff's alpha', whose metric accounts for the probability of agreeing by chance [22]. These evaluations, per article part, aggregate classification and for articles where annotations agree fully, 'all parts', are given in Table I. According to [22], a Krippendorff's alpha metric of $67\% \leq \kappa \leq 80\%$ can be used to 'draw tentative conclusions', whereas $\kappa < 67\%$ must be 'discarded'. Hence, agreement on all the considered article parts can be used to 'draw tentative conclusions', according to Krippendorff's alpha, except for the 'main' annotations, which must be 'discarded'. In our classification model we only try to predict the 'aggregate' classification, and will only use the 432 articles where agreement on 'all parts' were achieved as our dataset. We will now discuss discrepancies between 'aggregate' and 'all parts' annotations. Additionally, we will investigate any correlations between the annotated article parts, which can be exploited to enhance classification precision, if used as a heuristic in feature selection.

A. Discrepancies in 'Aggregate' Classifications

The annotators agreed on roughly two-thirds of the 'aggregate' sentiment classifications, as evident in Table I. Considering the confusion matrix, given in Table II, it is clear that most disagreements are between the *0* and *-1*, in addition to between the *0* and *+1* classifications. Furthermore, it seems that annotator 2 has a stronger polarity bias in that he uses *-1* and *+1* more frequently than annotator 1. This can be caused by annotator 2 being more familiar with the financial news domain and, hence, more

Table IV: Textual Feature Category (X)

Feature	Description
LengthOfTitle	Number of words in title
LengthOfLeadText	Number of words in lead text
LengthOfMainText	Number of words in main text
AverageLengthOfWords	Average character length of words
NumberOfExclamationMarks	Number of exclamation marks
NumberOfQuestionMarks	Number of question marks
NumberOfQuotes	Number of quotes

Table V: Categorical Feature Category (C)

Feature	Description
CategoryAnalysis	Whether published in Analysis category of hegnar.no
CategoryEconomics	Whether published in Economics category of hegnar.no
CategoryStockExchange	Whether published in Stock Exchange category of hegnar.no

Table VI: Grammatical Feature Category (G)

Feature	Description
NumberOfAdjectives	Number of adjectives
NumberOfNouns	Number of nouns
NumberOfVerbs	Number of verbs
NumberOfNegativeAdjectives	Number of adjectives with negative sentiment value
NumberOfPositiveAdjectives	Number of adjectives with positive sentiment value
NumberOfNeutralAdjectives	Number of adjectives with neutral sentiment value
NumberOfPositiveVerbs	Number of verbs with positive sentiment value
NumberOfNegativeVerbs	Number of verbs with negative sentiment value
NumberOfPositiveAdverbs	Number of adverbs with positive sentiment value
NumberOfNegativeAdverbs	Number of adverbs with negative sentiment value

Table VII: Contextual Feature Category (X)

Feature	Description
RecommenderCluesMentioned	Whether any recommender clue words occur in article
AnalyticsCluesMentioned	Whether any analytics clue words occur in article
HasPositiveTitleClues	Whether any positive clue words occurring in title
HasNegativeTitleClues	Whether any negative clue words occurring in title
PositiveTitleClueCount	Number of positive clue words occurring in title
NegativeTitleClueCount	Number of negative clue words occurring in title

often refraining from annotating with *0*, which was agreed to be used whenever in doubt. Additionally, annotator 2 spent slightly longer time annotating the articles. With this observation one could speculate to whether this has allowed him to read the articles more thoroughly and noticing sentiment-bearing expression more often.

B. Discrepancies in 'All Parts' Classifications

It might seem disquieting that the simple agreement 'all parts' is merely 43.2%, as can be seen in Table I. However, this agreement, as stated, necessitates agreement on all of the four annotations and, hence, agreement by chance is fairly low. This is embodied in the 'Krippendorff's alpha' statistic of 73.2%, the highest reliability of all the compared annotations. This reassures the selection of these 432 articles as the dataset to be used for further analysis.

C. Correlations Between Article Part Classifications

Considering the correlations between the different annotated parts (a matrix of correlations is given in Table III), a few interesting observations can be made. Not surprisingly, there are strong correlations between 'aggregate' classification, which is of our interest, and the other article parts. Indeed, variations in 'title' classification explains 88.6% of the variations in 'aggregate' classification. This is important to note because, first of all, one can then opt to only using the title in a quick-and-dirty classification of article aggregate sentiment. Secondly, one can capitalize on this correlation by selecting features from the article title as these are highly indicative of the aggregate sentiment.

IV. Feature Engineering

A selection of 26 features were extracted from each of the articles in the dataset. The feature selection was based on 1) prior efforts in the field and 2) hypotheses about the determinants of sentiment after closely examining a larger collection of articles on hegnar.no. For instance, we believed that question- and exclamation marks often to be associated with sentiments. Furthermore, we noticed that certain categories were more prone to sentiments than others (e.g analyst recommendations, economy). The relative over-representation of features directly related to the article title have been added due to the revealed strong correlation between title and aggregate sentiment, as discussed in the previous section. The set of 26 features were divided into four categories, representing the underlying context of the given feature, and will now be accounted for in turn. We have also included the formal definitions of the feature categories that we used to identify which category a feature belonged to.

A. Textual Features (T)

The Textual feature category, abbreviated T, contained each feature that had been derived only using the textual information available in the articles. See Table IV for an overview of these features with description and type. Formally:

Definition 1. *Textual features are extracted using only explicitly contained information in the text, like word count, character length of words, occurrences of quotes, exclamation and question marks.*

B. Categorical Features (C)

The Categorical feature category, abbreviated C, contained all the features that were created using the categories of the articles. These were extracted from the url of the article. See Table V for an overview of these features with description and type. Please note that the categories are non-exhaustive and mutually exclusive (i.e. each article can belong to none or only one category). Formally:

Definition 2. *Categorical features are extracted from the categorical information contained in the text, like the news publisher's own categorization tag set.*

Figure 2: Algorithm searching exhaustively for optimal feature subset using machine learning classifier c

Input: c: machine learning classifier, F: feature candidate set, D: article data set
Output: f^*: optimal feature set

```
maxPrecision ← 0
f* ← {∅}
for f ∈ 2^F do
    precision ← 0
    for i = 1 to 5 do
        trainingset ← EXTRACT-TRAINING(D, i, f);
        testset ← EXTRACT-TEST(D, i, f);
        c ← TRAIN(c, trainingset);
        precision ← precision + EVALUATE(c, testset);
    end for
    averageprecision ← precision / 5;
    if averageprecision > maxprecision then
        maxprecision ← averageprecision;
        f* ← f;
    end if
end for
return f*
```

C. Grammatical Features (G)

Features residing in the Grammatical category, abbreviated G, were created using grammatical information contained in the text of the news articles coupled with simple domain-independent sentiment lexica. A list of valence shifters (*ikke*, *aldri* etc. - *not*, *never*) was constructed. In addition, a simple domain-independent sentiment lexica was devised by associating sentiment classifications (*-1*, *0* or *+1*) with words commonly occurring on hegnar.no, within different word classes. To identify the different word classes, a part-of-speech (POS) tagger was used. When counting the occurrences of positive or negative words within the different classes adjustment for valence shifters was done (e.g. *ikke bra* - *not good* - would count as a negative adjective). See Table VI for an overview of these features with description and type. Formally:

Definition 3. *Grammatical features are extracted using the linguistic category of words (part of speech), valence shifters and simple domain-independent sentiment lexica. For instance, the phrase 'ikke bra' ('not good') can in a feature considering (valence shifted) verbs count as contributing negatively to the overall sentiment.*

D. Contextual Features (X)

The Contextual feature category, abbreviated X, includes all features that use domain-specific sentiment clue dictionaries. These were devised by manually associating sentiment classifications (*-1*, *0* or *+1*) with words thought to be indicative of a recommendation, an analysis, a positive and negative sentiment-bearing title. See Table VII for an overview of these features with description and type. Formally:

Definition 4. *Contextual features are extracted using domain-specific sentiment dictionaries, devised specifically for the document collection at hand. In the Norwegian financial news domain this includes adding sentiment entries for jargon and lingo typical of this type of news, like 'børsrakket' ('rocketing stock') or 'kontraktsdryss' ('awarding of numerous contracts').*

Table VIII: Best feature composition by classifier

Feature	Category	SVM	ANN	NB	RF	J48
NumberOfExclamationMarks	T					
NumberOfQuestionMarks	T	✓	✓		✓	✓
CategoryAnalysis	C	✓		✓		
CategoryStockExchange	C				✓	✓
NumberOfPositiveAdjectives	G					✓
NumberOfNegativeAdjectives	G					
RecommenderCluesMentioned	X	✓	✓		✓	
AnalyticsCluesMentioned	X					
HasPositiveTitleClues	X	✓	✓	✓		
HasNegativeTitleClues	X	✓	✓	✓		✓
PositiveTitleClueCount	X	✓	✓		✓	✓
NegativeTitleClueCount	X	✓	✓	✓	✓	✓

E. Feature Subset Search

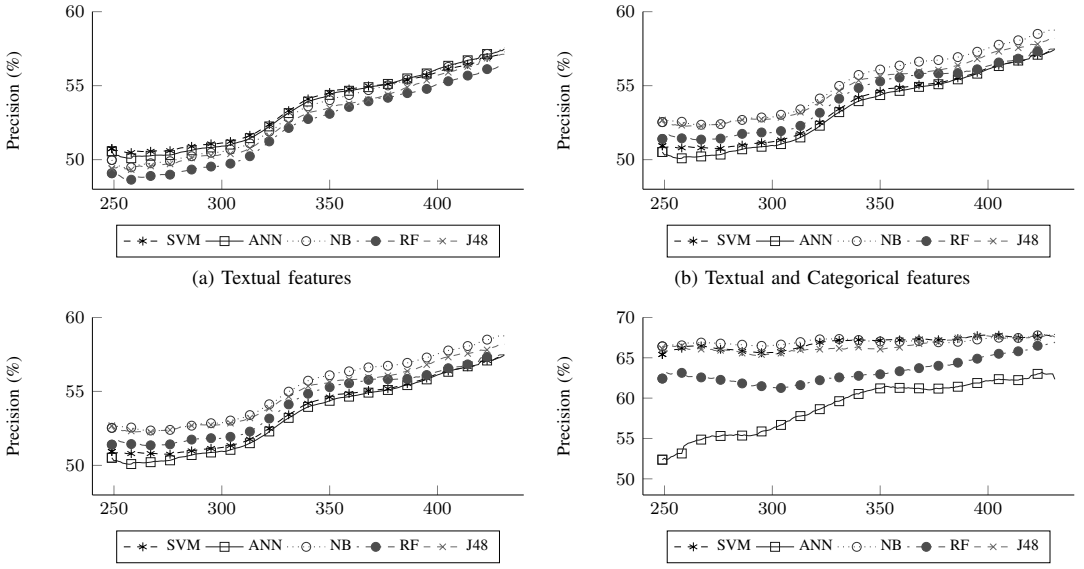
Not all of the identified and extracted features will contribute (equally) to the classification model. Given the size of the search over the feature subset space being $2^{26} = 67,108,864$, and the time needed to train and evaluate our classification models, exhaustive search was infeasible. We ranked the 26 features by Mutual information, χ^2 statistic and Maximal information coefficient [20] and used these as heuristics for selecting the 12 most promising features, and ran exhaustive searches over the subsets. The algorithm used in performing this exhaustive search is given in Figure 2. Here the method EXTRACT-TRAINING(D, i, f) returns a $trainingset$ of all but the i-th fold of features f in dataset D. EXTRACT-TEST(D, i, f) does the same but returns the $testset$ corresponding to the i-th fold. EVALUATE(c, $testset$) evaluates machine learning classifier c on $testset$.

These exhaustive searches resulted in 5 features in the optimal feature subset when evaluated with the classifiers Naïve Bayes (NB) and Random Forrest (RF), 6 features for Artificial Neural Networks (ANN) and J48 and 7 for Suppor Vector Machines (SVM). The details of the best (i.e. optimal) feature subset composition by classifier shown in Table VIII. We also experimented with various local search techniques over the full 26 features, but this did not result in superior results to the exhaustive subset search - confirming the validity of the statistics used to rank the initially selected 26 features.

V. Classification Training and Testing

We used five different machine learning methods for training and classification; Artificial Neural Networks (ANN), Support Vector Machines (SVM), Naïve Bayes

Figure 3: Classification precisions with different feature subsets and growing datasets (smoothed over 10 steps)



(a) Textual features

(b) Textual and Categorical features

(c) Textual, Categorical, and Grammatical features

(d) Textual, Categorical, Grammatical, and Contextual features

(NB), Random Forrest (RF), and J48 allowing for simple comparisons. The former three were chosen as they have both been widely used in previous related work. SVM is, in particular, commonly regarded as the highest performing classifiers in sentiment analysis [8]. RF and J48 were also included due to their recently revealed effectiveness in sentiment classification [5], [6].

ANN and SVM were set up using Encog (www.heatonresearch.com/encog) and NB, RF and J48 with the Weka framework (www.cs.waikato.ac.nz/ml/weka). Only minor parameter tuning was performed and this was done with all features included in the model. Hence, performance increases could be expected from fine-tuning the classifiers for each feature subset, however, this was not of emphasis in this paper. For SVM we used epsilon support vector regression with a radial basis kernel function. The ANN had four layers and the internal layers twice as many nodes as the input and output layers. For NB, RF and J48 less tuning was necessary and Weka's standard parameter values were used. We trained all classifiers using a random permutation of the training set as well as averaged 5-fold cross-validation.

Table IX: Classification Precision Results (%)

Features Categories	SVM	ANN	NB	RF	J48
T	57.4	57.7	**57.8**	56.7	57.3
TC	57.1	57.8	**59.1**	57.6	57.9
TCG	57.6	58.7	**59.0**	58.2	56.9
TCGX	66.5	64.0	68.2	66.9	**68.7**
BEST*	68.4	65.8	70.1	70.2	**70.8**

*Best feature composition

To test the effectiveness of our classifiers we ran tests using a selection of the most rewarding features from every feature category, as defined in section IV. The classification precision results are given in Table IX. These are listed in Table VIII. We also ran tests with growing datasets to see how the five classifiers performed at various dataset sizes - indicative of the value in attaining and employing a yet larger dataset.

All five classifiers performed quite well, compared to the state-of-the-art [2], [3], [4], when using features from all the four categories, in addition to those derived from the exhaustive searches, referred to as 'best feature composition'. The Textual feature category performed, unsurprisingly, the worst. After including features from the Categorical and Grammatical categories, we see only slight improvement for all classifiers. The inferior classification precision results of the three first feature categories could be a result of too little training data, or, perhaps more likely, simply that the features are simply inept determining sentiment in this context. The former explanation finds support in the graphs showing classifications with growing datasets, shown in figures 3a, 3b, and 3c. These graphs show a clear positive correlation between classification precision and dataset size, and no signs of convergence. Hence, we have reason to believe that a larger annotated dataset could increase the classification results even when just relying on these feature categories.

A major jump in classification performance was achieved once the features composing the Contextual category were included. The improved classification could be a result of these features being derived using domain specific knowledge, and, therefore, having a stronger impact on

the classification task. Another explanation could be that many of the features in the Contextual category reflect the characteristics of the article's title. The evident strong correlation between the title and aggregate classification, as accounted for in section III-C, explains the impact of features that are based on this article part.

The best results were achieved with a set of features that were derived from the exhaustive searches, performed independently for each of the five classifiers, as detailed in section IV-E. These sets were only slightly different; all subsets were dominated by features from the Contextual category, as evident in Table VIII. This was as expected since this category contained the most rewarding features judging by the different feature rankings (Mutual Information, χ^2 and Maximal Information). The distinctions between the five classifiers can be explained by the machine learning methods naturally benefiting from different combinations of features.

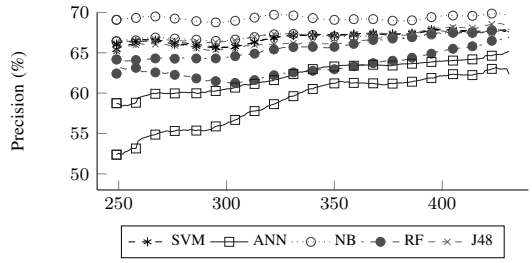
Considering the 'best feature composition' (BEST), J48 outperforms the other four, which is in line with the most recent reviewed literature [5], [6]. The same also holds when all the four different feature categories are included (TCGX in Table IX). This outperformance is only marginal compared to NB and RF for these two feature subsets. According to its creator, RF has unexcelled accuracy compared to most classifiers [23]. J48's slightly higher accuracy aligns with it having been found to perform better for a relatively smaller number of instances [7], as we're working with. One could expect RF to perform better if attaining yet a larger dataset. It is interesting to note that NB, RF and J48 achieve notably higher accuracies than SVM and NB for the two aforementioned feature subsets. These former classifiers have less need, and room, for parameter tuning than the two latter. Hence, the evident performance divergence could be explained by tuning being de-emphasized in this paper.

For the three other feature subsets (T, TC and TCG) NB achieves the highest precision. This classifier performing better in comparison to the others for these feature subsets can be explained by the conditional independence assumption being less violated when working with lower dimensionalities [24], which is the case for these categories.

The discrepancies in SVM and ANN performance deserves a discussion since 1) they have both been widely used in previous work and 2) they are comparable in necessitating relatively more parameter tuning then the other three classifiers. In our case, these variances can be explained by SVM's ability of always finding a solution that is both globally optimal and unique, whereas ANN tends to get stuck in local maxima. This explanation is probable given the feature space landscape being rich of local optima, as noted in section IV-E. Other possible explanations for these discrepancies could be that we only permit a limiting number of training iterations for the ANN; out of run-time considerations. It is also known that SVM is less prone to overfitting, and this could also be reflected in the classification results.

J48's superior precision is fortunate, considering potential real-time employment of the sentiment engine, since it has low running time both in training and classification - especially compared to SVM and ANN. Running time requirements may not necessarily be an issue in classification, but if our system needs to be retrained periodically using a slower classifiers, like ANN, can be a constraint.

Figure 4: Classification precision with 'best feature composition' and growing dataset



The best achieved precision of ∼71% is comparable to previous efforts, as described in II, and clearly exceeds the most similar previous study (news domain) having obtained a precision of ∼54% [1]. We have yet to derive a human classification baseline, based on our annotation guide. However, one could argue that the achieved annotator agreement, on the aggregate sentiment classification, of 73.4% could serve as a proxy for this. In that sense, our best achieved precision result indicate that the performance of the model is close to that of a human.

VI. Conclusions and Further Work

We have proposed, defined, experimented with, and evaluated, different feature categories and classifiers in sentiment analysis of Norwegian financial news articles. Using machine learning techniques, with 26 article features grouped in four categories as input, we achieve a precision of ∼71%, comparable to the state-of-the-art in other domains and close to our human baseline. In our analysis, we have found features relying on context-specific sentiment lexica ('contextual' category) to be paramount in classification within this domain - yielding a precision increase of ∼21% when added to the other feature categories. Furthermore, we have found J48 classification trees to yield the highest classification performance, closely followed by Random Forrest (RF), in line with recent studies [5], [6], [7], and in opposition to the antedated conception that Support Vector Machines (SVM) is superior in this domain [8].

Additionally, and although there are signs of modest precision convergence, there are still indications of enhanced results with a larger dataset. Furthermore, the incremental increase precision from adding grammatical features, like valence shifters, holds promise of further enhanced results if extending the set of features by drawing on further linguistic theory. Accounting for n-grams, co-occurring terms [25], developing and using a more refined part-of-speech (POS) tagger, in addition to performing

stemming include some of the linguistic extensions of interest going forward. If these venues of further work are properly investigated, leading to sufficient sentiment classification enhancement, interesting, remunerative applications can follow from the study being focused on the psychology-rich Norwegian financial domain.

ACKNOWLEDGEMENTS

The authors would like to thank Jon Espen Ingvaldsen and Arne Dag Fidjestøl for technical help in addition to three anonymous reviewers. This work is partially funded by NxtMedia (www.nxtmedia.no) and the Telenor Group (www.telenor.com).

REFERENCES

[1] A. Balahur, R. Steinberger, M. A. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news." in *LREC*, 2010.

[2] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 70–77.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[4] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[5] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[6] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

[7] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees." *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, 2012.

[8] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012, pp. 90–94.

[9] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[10] Ø. Gjerde and F. Saettem, "Causal relations among stock returns and macroeconomic variables in a small, open economy," *Journal of International Financial Markets, Institutions and Money*, vol. 9, no. 1, pp. 61–74, 1999.

[11] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.

[13] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Computing attitude and affect in text: Theory and applications*. Springer, 2006, pp. 1–10.

[14] F. Simančík and M. Lee, "A ccg-based system for valence shifting for sentiment analysis," *Research in Computing Science*, vol. 41, pp. 99–108, 2009.

[15] A. Sharma and S. Dey, "A document-level sentiment analysis approach using artificial neural network and sentiment lexicons," *ACM SIGAPP Applied Computing Review*, vol. 12, no. 4, pp. 67–75, 2012.

[16] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature subsumption for opinion analysis," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 440–448.

[17] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 3, pp. 447–462, 2011.

[18] Z. Zhai, H. Xu, J. Li, and P. Jia, "Feature subsumption for sentiment classification in multiple languages," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2010, pp. 261–271.

[19] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 12, 2008.

[20] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[21] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, vol. 6, 2006, pp. 417–422.

[22] K. Krippendorff, "Reliability in content analysis," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naïve bayes vs. decision trees vs. neural networks in the classification of training web pages." *International Journal of Computer Science Issues (IJCSI)*, vol. 7, no. 4, 2010.

[25] Y. Tao and J. X. Yu, "Finding frequent co-occurring terms in relational keyword search," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, 2009, pp. 839–850.

Chapter 7

Paper II

Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, and Jon Atle Gulla: *Optimizing Supervised Sentiment Lexicon Acquisition: Selecting Co-Occurring Terms to Annotate for Sentiment Analysis of Financial News*, submitted to Springer's Language Resources and Evaluation.

(This page is intentionally left blank.)

Optimizing Supervised Sentiment Lexicon Acquisition: Selecting Co-Occurring Terms to Annotate for Sentiment Analysis of Financial News

Pål-Christian Salvesen Njølstad · Lars Smørås Høysæter · Jon Atle Gulla

Abstract A common supervised approach to sentiment analysis is counting occurrences of sentiment-bearing Co-Occurring Terms (COTs) in texts and looking these up in a sentiment lexicon. These occurrences are then aggregated and used as input features to a machine learning model which is trained and tested on an annotated text corpus. When faced with a new corpus in a different domain, however, this approach requires new lexica to be derived because different domains 1) use diverging vocabularies of words, requiring new lexicon entries, and 2) have disparate sentimental interpretations of the same words. Most research on lexicon acquisition requires the availability of extensive lexical resources as it focuses on either automatically extending pre-existing lexical resources or inferring contextual lexica from texts. The conjecture is that manual lexicon acquisition is infeasible given the effort required to achieve satisfactory sentiment classification precision. In this paper we impugn this supposition and present a novel proof-of-concept system affirming that state-of-the art sentiment classification precision can be achieved in a new and different domain (Financial news written in Norwegian) in absence of comprehensive lexical resources and with permissible manual annotation efforts.

Furthermore, this paper investigates the number of COTs that needs to be annotated, which radius (permitted number of words between the COTs) should be used, and how the COTs candidate for entry in a sentiment lexicon should be ranked in order to achieve satisfactory precision. Our experimentation finds the ranking function mutual information to be superior, that precision increases diminishingly with lexica size and weakens with larger radii.

Keywords Sentiment Analysis · Lexicon Acquisition · Manual Annotation · Machine Learning · Financial News

P. C. S. Njølstad (✉)
Department of Computer and Information Science
Norwegian University of Science and Technology
E-mail: palchrnj@stud.ntnu.no

L. S. Høysæther
E-mail: larssmor@stud.ntnu.no

J. A. Gulla
E-mail: jag@idi.ntnu.no

1 Introduction

With the advent of the internet the amount of available information has grown exponentially (Huberman and Adamic, 1999). This panoply of available information, generated by a multitude of users, represents a valuable source of opinions for decision-makers. In an increasingly fast-paced and competitive world, an organization's ability to effectively make sense out of opinions expressed in internet documents might turn out to be key in maintaining competitiveness or fulfilling its purpose going forward. More specifically, a company manufacturing products could keep the opinionated pulse of its existing and prospective customers as it considers which new products to develop (Yi et al., 2003). A mutual fund holding a diversified portfolio of stocks could monitor the sentimental landscape associated with its current assets opined by journalists of various financial news papers (Parikh et al., 2012). Or a hedge fund could process sentiments from news stories, filings, social media and blogs real-time and make trades in anticipation of stock price behaviour (Andrews et al., 2011), which, according to Feldman (2013), can lead to superior returns.

Undeniably, as the number of opinionated documents of interest grows it becomes practically impossible to manually analyse these in an effective manner. Hence, research addressing the scalability of this problem has emerged, and soon evolved, drawing on theory from the intersection of computational linguistics and information retrieval (Wei and Gulla, 2010). Early influential work includes that of Hu and Liu (2004), Pang and Lee (2004), Pang et al. (2002), Hatzivassiloglou and McKeown (1997) in addition to Dave et al. (2003). Recently, Socher et al. (2013) claimed to augment the single-sentence state-of-the-art product review sentiment classification precision from ∼80% to 85.4%. In terms of content, document domains composed of news articles (Balahur et al., 2013), social media text snippets (Pang and Lee, 2008), blog posts (Chesley et al., 2006), and political speeches (Yu et al., 2008), among others, have been subject to extensive research.

Independent of application domain, sentiment lexica usually play a central role in performing the required classification task (Lu et al., 2011). According to Feldman (2013), sentiment lexicon acquisition, which is the focus of this paper, remains as one of the main specific problems within sentiment analysis. The consensus within this research field is that domain-independent universal sentiment lexica are futile (Qiu et al., 2009; Lu et al., 2011; Turney, 2002). Wilson et al. (2005) quantify this inferiority: they found a simple classifier and a universal lexicon to yield a precision of 48% compared to 65.7% when using a contextual lexicon.

As such, much effort has been directed towards automatically extracting lexicon entries with classification from a collection in new domains (Godbole et al., 2007; Tan et al., 2008; Lu et al., 2011) in addition to automatically extending sentiment lexica from one domain or application to others (Qiu et al., 2009; Neviarouskaya et al., 2009). These approaches require having either a fairly large dataset in order to use statistical methods for entry extraction or the availability of certain lexical resources, e.g. a sentiment lexicon and / or a synonym dictionary. Within some application domains, however, neither of these requirements can easily be met. For instance, corpus size might be small if seeking to perform sentiment analysis of the news flow associated with a particular stock over a limited period of time. This is especially true if the stock is of limited public interest, e.g. being less traded, less covered by analysts and / or listed on a smaller exchange, like the Oslo

Stock Exchange (OSE). These stocks, being less efficient, are according to works on value investing, like that of Graham et al. (1934) and Graham (1959), exactly the stocks one should seek to analyse. Additionally, and in the specific case of stocks listed on OSE, the associated news flow will be in Norwegian - a language that, like most languages but English, has meagre lexical resources, as publicly available sentiment lexica and synonym dictionaries are non-existent (Perez-Rosas et al., 2012). This lack of resources can, if able to develop an ample sentiment classification engine, potentially give a remunerative competitive advantage over other traders or investors on the OSE.

Hence, for such a domain, which is subject to our particular study, selected because of its demonstrative as well as potentially pecuniary value, neither automatic sentiment lexicon with classification extraction nor lexicon extension are viable approaches. In lieu of these techniques, acquisition of a sentiment lexicon in a supervised fashion represents a feasible option. This is, as some will argue, less desirable but appears to be the best alternative given the aforementioned constraints. What then needs to be investigated is then 1) how much effort it requires and 2) how one best carries out such a procedure in achieving satisfactory sentiment classification performance. This we do by optimizing supervised extraction of COTs for entries in a sentiment lexicon used for sentiment analysis in the financial news domain.

The main contributions of this paper are two-fold. Firstly, we present and evaluate a sentiment classification system that, with fairly low annotation efforts, achieve satisfactory document-level sentiment classifier when presented with a new corpus in domain with limited size and no pre-existing suited sentiment lexicon. Our achieved classification precision results are comparable to state-of-the-art work in the news article domain, such as (Balahur et al., 2010). This approach having been deemed infeasible in previously published literature, like in Das and Chen (2007) and Feldman (2013), confirms the novelty of our contribution. Secondly, we optimize the process of deriving this sentiment lexicon by experimenting with and evaluating five different COT ranking functions, three different lexica sizes, varying radii in addition to three different machine learning classifiers.

We will now outline the remainder of this paper. An account of the work related to our paper is given in section 2. Followingly, the two main contributing parts of our research are accounted for in the two subsequent sections:

- *System Overview* - We develop and evaluate a system for performing supervised document-level sentiment analysis in the financial news flow associated with stocks on OSE. The details of this system, with justifications for design choices made, are given in Section 3.
- *Experiments and Evaluation* - We experimented with and evaluate the different parameters of our system - the input ranking function (f), the radius (r), sentiment lexicon size (σ, ρ) in addition to machine learning classifier (c). The classification results from varying these four parameters are accounted for and discussed in Section 4.

Finally, we conclude and briefly discuss interesting venues for future work, in section 5.

2 Related Work

Sentiment analysis, or opinion mining as it often is referred to, is performed either at the word, phrase, sentence, paragraph or document level and is generally formulated as a ternary classification problem; one seeks to classify analysed texts as either positive, negative or neutral (or on a multi-way scale, such as in (Snyder and Barzilay, 2007; Pang and Lee, 2005)). In this paper we focus on attaining the aggregate sentiment of articles and, hence, this related work section will focus on previous efforts conducted at the document-level. At this level of analysis one typically distinguishes between supervised and unsupervised approaches. The former assumes that there is a finite set of classification categories and that training data is available for all of these classes. This approach uses the training data to learn a classification model by employing the common machine learning algorithms, like Support Vector Machines (SVM) and Naïve Bayes (NB) (Annett and Kondrak, 2008). The latter approach determines the semantic orientation of document phrases and classifies the overall document-sentiment as positive (negative) if the average semantic orientation of its constituent phrases is above (below) a certain threshold. In this approach a POS tagger and / or a sentiment lexicon is generally employed (Feldman, 2013).

As previously noted, sentiment lexica are paramount in sentiment analysis. Feldman (2013) deems the acquisition of such lexica to be one of the main research problems in sentiment analysis and categorizes three main approaches to doing this: dictionary-based, corpus-based in addition to manual annotation approaches. These approaches are often used in conjunction. The dictionary-based approach typically starts with a set of domain-specific seed words which are then expanded using the antonymous and / or synonymous relations in a linguistic ontology, like WordNet (Miller, 1995). Kamps et al. (2004) did this by counting the number of edges between two words in a synonym word graph, Dragut et al. (2010) using devised deduction rules, and Peng and Park (2004) by comparing and reasoning about the overlap of word synonym sets. This approach has the drawback in that it generally does not result in domain-specific lexica.

Corpus-based approaches, on the other hand, address this limitation by being based on corpora, composed of documents in the domain one seeks to analyse, and extend a sentiment lexicon using a set of seeds. Hatzivassiloglou and McKeown (1997) did this by using linguistic connectors (*and*, *or* etc.), and the assumption of sentence coherency, to extend the polarity classification of known adjectives to classify newly encountered ones. A similar approach has been done based on co-location of words with the seed words *excellent* and *poor* (Turney, 2002). Kanayama and Nasukawa (2006) used successive appearance of known and newly encountered polar atoms, defined as the minimum human-understandable syntactic structures that specify the polarity of clauses, within the same context to attain new sentiment lexicon entries. A double propagation approach was used by Qiu et al. (2009), specific for the product review domain, which leverage the co-occurrences between feature and sentiment words to learn both new features and sentiment lexicon entries. Although corpus-based approaches can help find domain-specific opinionated words and their orientations, they are limited in that it is difficult to attain a dictionary adequately covering the verbosity of a language (Liu, 2010).

Manual annotation, as an approach to acquiring lexica, is the process of having one or (preferably) more humans going through words and classifying the sentiment of these in composition of a dictionary. Some of the fixed and manually devised publicly available resources include Bing Liu's Opinion Lexicon (Liu, 2010), the Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon (Wiebe et al., 2005), SentiWordNet (Baccianella et al., 2010), the Harvard General Inquirer (Stone et al., 1966), and the Linguistic Inquiry and Word Counts (LIWC) database (Pennebaker et al., 2001). The approach of manual annotation is, according to Liu (2010), not usually used alone but combined with automated approaches as a final check since the latter often commits errors. Using the manual annotation process alone is infeasible, according to Feldman (2013), given its labour-intensity (Das and Chen, 2007) and the assumption that each each domain requires its own lexicon.

The difficulty in carrying out annotation tasks in the realm of sentiment analysis, such as constructing sentiment lexica, sheds light on the imprecise nature of sentiment classification. In (Rubin, 2007) an agreement of ∼71% was achieved among annotators on quinary classification. The author ascertains that independent annotators' subjective perceptions of annotation class boundaries present great difficulty in manual annotation. Njølstad et al. (2014) achieved agreement between 63% and 75% in annotating different article parts (title, lead and main text) and in (Hsueh et al., 2009) a ∼77% agreement in ternary annotation of snippets of text was attained. Furthermore, Wilson et al. (2005) found a ∼82% agreement in the assignment of phrase-level sentiment polarity, however this study only included sentiment bearing documents and, hence, concerns the simpler binary classification task. The precision of a sentiment analysis system is measured by how well it agrees with human judgement and, as such, the ability of humans to agree among themselves must be an upper bound to the precision we can realistically expect from such a system. From the agreement statistics presented, ranging from ∼73% to ∼82%, slightly diverging due to the text and the arity of the classification task considered, a system exhibiting a ∼70% accuracy is actually performing quite well, close to that of a human, even though the figure in itself may not sound all that impressive.

If inducing a lexicon directly from data, which in absence of ample lexical resources in practice only can be done by the means of manual annotation, one can advantageously capture domain-specific effects, which has the propensity to determine presence and polarity of sentiment (Liu, 2010). The questions that then remain are how much effort is required to achieve a satisfactory classification precision and how one best goes about doing this. According to Perez-Rosas et al. (2012) only a small number of lexica for non-English languages, which generally have scarcer lexical resources, have been manually devised. This includes Abdul-Mageed et al. (2011) who compiled and annotated a list of 4,000 Arabic adjectives from the newswire domain and Clematide and Klenner (2010) who extracted a list of 8,000 nouns, verbs, and adjectives in German and annotated these with both polarity and strength. In terms of effort, the number of data points are in short supply. Wiebe et al. (2005) report having had two annotators spend 8-12 hours per week for 3-6 months in annotation of a 10,000 sentence corpus (a more elaborate task then annotating simple words or COTs). And Clematide and Klenner (2010) used a setup where each adjective had to be annotated in a time frame of at most 12 seconds - which when constructing a large lexicon will amount to a significant

amount of time. In general, previous literature use other lexical resources to guide their selection of candidates for annotation and entry in the lexica. Clematide and Klenner (2010), for instance, use GermaNet, a WordNet-like lexical database, as a heuristic. This, however, resulted in a stock of adjectives insufficient for sentiment analysis of German novels. In (Abdul-Mageed et al., 2011) it is unclear how the adjectives were selected, other than that they 'pertain to the newswire domain'. This lack of a guided procedure for selecting candidates for entry and annotation in sentiment lexicon, especially in absence of other lexical resources serving as a heuristic, justifies such an investigation. Moreover, most studies of manual lexicon acquisition, like that of Abdul-Mageed et al. (2011), Clematide and Klenner (2010) and Wiebe et al. (2005), evaluate their lexica by comparing annotations between either annotators and / or other lexica, and not by applying it in performing actual sentiment analysis. This previous decoupling between lexicon acquisition and end-task evaluation further justifies our investigation.

The previous work most akin to our coupling of lexicon acquisition and end-task evaluation is Mourad and Darwish (2013), who use the manually constructed lexica detailed in (Abdul-Mageed et al., 2011), ArabSenti, to perform sentiment analysis of Arabic news and tweets. Since this lexicon was evaluated both with and without extensions, their work can serve as a basis of comparison. The authors report classification precisions results ranging from 64.8% to 80.4%, depending on the domain and whether a subjectivity or polarity classifier was used. Specifically, the authors achieved a precision of 71.1% on polarity classification of tweets and 80.4% on single sentences form the news domain. These results, however, are not directly comparable to our work since it is inherently more difficult to perform 1) ternary than binary classification and 2) document-level sentiment analysis, as oppose to that of single sentences or tweets, due to the increased complexity in having to aggregate the, perhaps sentimentally contradicting, multitude of sentences composing, in our case, an article. Arguably, this is confirmed by Mourad and Darwish (2013) achieving higher precision for single sentences than tweets, which on average are shorter than the 140 signs allowed to be twittered (the average Arabic sentence length is 17.4 words (Dorr et al., 2002) and the average character word length is 5 (Alotaiby et al., 2009)). Hence, we advocate the precisions towards the lower part of Mourad and Darwish's (2013) precision range, like the 71.1% for tweets, to be the best comparable to our investigation.

Having constructed a sentiment lexica a few vital steps need to be carried out before arriving at a trained machine learning classifier for the sentiment analysis of newly encountered documents. Firstly, each of the documents in the data set need to be converted into a numerical vector, which is what the machine learning algorithms operate on, a process known as feature extraction (Manning et al., 2008).

In the most common approach feature extraction approach, called Bag of Words (Whitelaw et al., 2005), unigrams (single term) or COTs (two or more words co-occurring in a sentence) form the basis of feature vectors (Annett and Kondrak, 2008). Co-occurring terms have the advantage over unigrams that they have the ability to capture some of the semantic, lexical, or other relations between terms (Matsuo and Ishizuka, 2004). When unigrams og COTs are combined with a sentiment lexica, feature vector elements can be created by counting the number of occurrences of different lexica entries in different parts of the text, for instance the number of positive COTs occurring in the document's title. Addi-

tional features that can be extracted or used to modify elements in the feature vector include valence shifters (negations and other words modifying the polarity being opined), but-clauses, purely textual features (document length, length of title, etc.), categorical features (if able to categorize documents based on content), grammatical features (if lexical resources, like a POS-tagger, are readily available) among others (Liu, 2010).

Secondly, the machine learning classifier needs to be trained and evaluated on the training and test data, respectively. Support Vector Machines (SVM) and Naïve Bayes (NB) are two of the most commonly used classifiers in sentiment analysis (Annett and Kondrak, 2008). The erstwhile claimed primacy of the former classifier (Wang and Manning, 2012), has recently been tested by J48 and Random Forrest (RF) having been found to yield the highest performance in this domain (Thelwall et al., 2010; Castillo et al., 2011; Ali et al., 2012).

3 System Overview

In this section we will state our formal definition of COTs, the rules governing the extracting of these from texts, along with illustrative examples, and briefly discuss additional constraints added to the formal definition in candidate COTs generation. Additionally, we detail the developed system, parameters to be optimized in lexica acquisition and design choices we have made for performing the supervised document-level sentiment analysis.

3.1 Co-Occurring Terms (COTs)

In our investigation we use COTs as the basis of sentiment lexica as they advantageously, as oppose to unigrams, have the ability to capture semantic, lexical, or other relations between terms (Matsuo and Ishizuka, 2004). We state our formal definition of COTs as follows.

Definition 1 (Co-Occurring Terms) Co-Occuring Terms (abbreviated COTs) are terms that co-occur in the same text without being separated by any of the punctuations: periods, question marks, or exclamation marks.

This definition is marginally different from the one used in other literature, like the one used by Wei et al. (2013). These authors also require COTs to be unseparated by the punctuations commas, colons and semicolons. We relax this requirement in our definition since the Norwegian language, to a greater extend than the English language studied by Wei et al. (2013), uses punctuations, and especially commas, unsparingly. A manifestation of this phenomenon is the first rule in the set of ten commandments for the proper use of Norwegian: 'It is not a shame to write a dot' (Santos, 1998).

Definition 1 implies that two terms occurring in the same document are not necessarily COTs, if separated by named punctuation. In the article title *Eier aksjer verdt 180 millioner. Kona får ny sykkel* (*Holds stocks worth 180 millions. Wife gets new bike*) the terms *aksjer* (*stocks*) and *kona* (*wife*) are not COTs being separated by '.', whereas the terms *aksjer* (*stocks*) and *verdt* (*worth*) are COTs. COTs are closely related to n-grams but more generic in that contiguous

appearances of words in text are not required. In the same sentence *aksjer* (*stocks*) and *millioner* (*millions*) are COTs but not an n-gram since the terms do not appear contiguously.

It is usual to add additional constraints when extracting COTs from documents, reducing the vast candidate space to consider. This is done by 1) confining the arity of COTs and 2) limiting the COTs radius, each of which we now will define and provide examples of.

Definition 2 (Arity) The arity of Co-Occurring Terms (COTs) is the number of terms they are composed of.

In the above sentence the COTs *Kona får sykkel* (*Wife gets bike*) have arity 3 whereas *Kona sykkel* (*Wife bike*) have arity 2.

Definition 3 (Radius) The radius of Co-Occuring Terms is the maximum allowed distance between the terms that are the furthest apart. The distance is the number of words between and including these outermost terms.

In the same sentence, the radius of the COTs *Eier aksjer* (*Holds stocks*) is 2, the radius of the COTs *Eier verdt* (*Holds worth*) is 3, and *Eier millioner* (*Holds millions*) is 5.

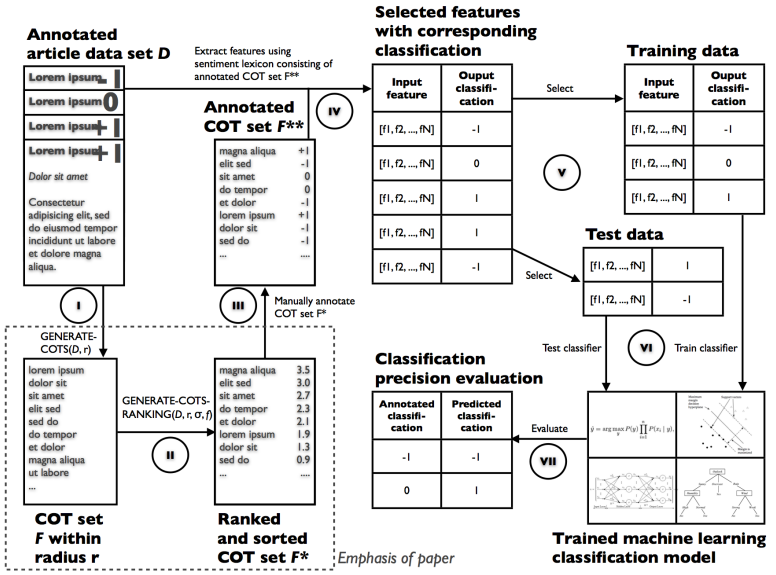
In our investigation we will only consider binary co-occurring terms out of simplicity and since Fürnkranz (1998) showed that n-grams (synonymous to COTs) with $n > 3$ contribute negatively to the slightly more general task of text classification. This means that all COTs will be composed of exactly two terms - *aksjer verdt* (*stocks worth*) is permitted but not *aksjer verdt millioner* (*stocks worth millions*). Furthermore, we will be varying the radius between COTs' constituent terms between 2 and 8. Under the assumption that the relationship between terms (be it semantic, lexical, or other) weakens as the distance between the terms increases and observing that few sentences in our domain are long we have capped the radius at 8. This is further elaborated on in Section 3.5.

In our investigation we further refine which word classes to examine, which we strictly limit to verbs, nouns, adjective and adverbs - the classes most prone to carry sentiments (Benamara et al., 2007). We have not reduced terms to the roots of its inflections in lack of lexical tools performing this task. We do, however, perform stemming to avoid entering and annotating different inflected variants of stems.

3.2 High-Level Description

A high-level illustration of the system is depicted in Figure 1. Input to the system is an *Annotated article data set D*. In order to generate a sentiment lexicon σ COTs within a radius r are extracted from the collection (step I), called *COT set F within radius r*. These are then ranked according to ranking function f (step II), arriving at a *Ranked and sorted COT set F**, and then manually annotated (step III), yielding an *Annotated COT set F***. The algorithms used for generating and ranking COTs are detailed in Section 3.3 and the ranking functions employed are described in Section 3.4.1. Accounts of the values selected for input parameters radius r and lexicon size σ are given in Sections 3.5 and 3.5.1, respectively.

Fig. 1: High-level illustration of system for sentiment classification. Emphasis of paper indicated in grey dotted box.



This annotated COT set then serves as a sentiment lexicon used in feature extraction: the occurrences of positive, neutral and negative COTs in the articles title and lead text are counted and form machine learning input features (Step IV). These features are detailed in Section 3.6. The *Selected features with corresponding classification* are then split into *Training data* and *Test data* (Step V). This training set is used to construct a *Trained machine learning classification model* (Step VI). The machine learning classification models employed are presented in Section 3.5.2. Lastly, we evaluate the machine learning model using the test set and comparing the predicted and annotated classifications (Step VII), arriving at the *Classification precision evaluation*. The evaluation results are accounted for and discussed in Section 4.

3.3 Algorithms

The emphasis of this paper is the derivation of the COT sentiment lexica later used as input to the feature extraction step (IV). This part of the system is enclosed within the grey dotted box shown in Figure 1. This part of the system includes both the COT generation and ranking, named GENERATE-COTS(D, r) and GENERATE-COTS-RANKING(D, r, σ, f) in the Figure. The algorithms performing these tasks will now be accounted for in turn.

Fig. 2: GENERATE-COTS(D, r): Algorithm for generating set of COTs in article data set D within radius r

Input: D: article data set, r: radius
Output: F_{cot}: set of COTs
1: $F_{cot} \leftarrow \{\emptyset\}$
2: **for** $d \in D$ **do**
3: $d \leftarrow$ CLEAN(d)
4: **for** $i = 1$ to LENGTH(d) **do**
5: $j \leftarrow 1$
6: $w \leftarrow$ EXTRACT-WORD-AT(d, i)
7: **if not** PERMITTED-POS(w) **then** {Only create COTs from permitted POS}
8: **continue**
9: **end if**
10: **while** $j \leq r$ and $v \notin \{.,!,?\}$ **do**
11: $v \leftarrow$ EXTRACT-WORD-AT(d, $i + j$)
12: $j \leftarrow j + 1$
13: **if not** PERMITTED-POS(v) **then** {Only create COTs from permitted POS}
14: **continue**
15: **end if**
16: **if** $((w \bigoplus v) \notin F_{cot}$ **then**
17: $F_{cot} \leftarrow F_{cot} \bigcup (w \bigoplus v)$
18: **end if**
19: **end while**
20: **end for**
21: **end for**
22: **return** F_{cot}

3.3.1 GENERATE-COTS Algorithm

The algorithm for generating candidate COTs, GENERATE-COTS(D, r), takes an annotated article data set D and a radius r as input and outputs a set of COTs which are candidates for annotation, F_{cot}. This algorithm is detailed in Figure 2. At a high level, this algorithm goes through the set of documents to be analysed, extracts all binary COTs within the specified radius and permitted word classes before returning them upon termination.

The function CLEAN(d) takes an article document d as input, tokenizes the text and returns all the words in the article along with sentence-ending symbols (., ! and ?) and all other symbols removed (,, :, ;, etc.). The PERMITTED-POS(w) function returns true if the word w belongs to the word classes noun, adjective, verb or adverb, and false otherwise. This is added since we are only interested in COTs where both terms are from one of these word classes. The function EXTRACT-WORD-AT(d, i) returns word at position i in article document d.

3.3.2 GENERATE-COTS-RANKING Algorithm

The algorithm for ranking candidate COTs, GENERATE-COTS-RANKING(D, r, σ, f), takes an annotated article data set D, a radius r, the number of COTs to be returned σ, and ranking function f as input and outputs a ranked set of COTs to be annotated, F^*_{cot}. This algorithm is presented in Figure 3. At a high level, this algorithm goes through the set of candidate COTs, ranks them according to

Fig. 3: GENERATE-COTS-RANKING(D, r, σ, f): Algorithm for generating ranked list of σ COTs based on article data set D within radius r and ranking function f

Input: D: article data set, r: radius, σ: length of ranked list to be returned, f: ranking function
Output: F^*_{cot}: ranked list of COTs
1: $F^*_{cot} \leftarrow \{\emptyset\}$
2: $F_{cot} \leftarrow$ GENERATE-COTS(D, r)
3: **for** $(w \bigoplus v) \in F_{cot}$ **do**
4: $tf_w \leftarrow$ TERM-FREQ(D, w)
5: $df_w \leftarrow$ DOC-FREQ(D, w)
6: $tf_v \leftarrow$ TERM-FREQ(D, v)
7: $df_v \leftarrow$ DOC-FREQ(D, v)
8: $tf_{w \bigoplus v} \leftarrow$ TERM-FREQ(D, $w \bigoplus v$)
9: $df_{w \bigoplus v} \leftarrow$ DOC-FREQ(D, $w \bigoplus v$)
10: **if** $df_{w \bigoplus v} \leq 1$ **then** {Require document frequency to be more than 1}
11: **continue**
12: **end if**
13: $F_{w \bigoplus v} \leftarrow$ COMPUTE-STAT(f, tf_w, df_w, tf_v, df_v, $tf_{w \bigoplus v}$, $df_{w \bigoplus v}$)
14: **end for**
15: $F^*_{cot} \leftarrow$ SORT-DESC(F_{cot}, F)
16: $F^*_{cot} \leftarrow F^*_{cot}(1 : \sigma)$
17: **return** F^*_{cot}

a ranking function (detailed in Section 3.4.1) and returns the top σ COTs in this ranked list. The output of this algorithm will then be subject to manual annotation in our system.

The algorithm gets the list of (unranked) candidate COTs by invoking the function GENERATE-COTS(D, r), which has been described in the previous section. In order to perform the ranking term and document frequency statistics need to computed per COT and constituent term. The function TERM-FREQ(d, w) returns the term frequency (i.e. number of occurrences) of the COT or term w in article data set D and DOC-FREQ(d, w) does the same for document frequency (i.e. number of documents with occurrences - multiple occurrences within the same document are not counted). Next, the function COMPUTE-STAT(f, tf_w, df_w, tf_v, df_v, $tf_{w \bigoplus v}$, $df_{w \bigoplus v}$) uses the ranking function f and the other input frequency variables to compute a ranking statistic for the COT $w \bigoplus v$. The ranking function f employs one of equations (1) - (5) to compute the statistic. The function SORT-DESC(F_{cot}, F) simply returns F_{cot} sorted descendingly based on the ranking statistic per COTs in F. Make note that we require COTs to occur in more than one article to be returned and, hence, annotated. The reason for this is the limited benefit of annotating COTs that only occurring in one article.

3.4 Parameters

We will now present, detail and discuss the parameters input to our system: the ranking functions (f), radius (r), lexicon size (σ) and machine learning classifier (c). These parameters will in turn be optimized and the results of this experimentation is later accounted for in Section 4.

3.4.1 Ranking Functions (f)

The number of candidate COTs in a corpus can get very large, especially when the radius r and the number of documents in the article data set D input to GENERATE-COTS(D, r) get sizeable. Although we have limited the COTs to only those composed of noun, adjective, verb and adverb terms, in addition to only COTs with a document frequency greater than 1, the number of candidates is still greater than what is reasonable to manually annotate. Hence, we need to find ways of ranking the candidate COTs and only annotate the top σ ranked candidates such that those suited for entry in a sentiment lexicon are included and those unsuited are left out.

In this work we are only looking at term and document frequencies of COTs and constituent terms in producing the rankings. When having an annotated article data set available, one approach could be using sentiment classifications in guiding the ranking of candidate COTs, however, this would require a much larger annotated data set - the same data set could not be used for building the lexicon and performing the classification without running the risk of overfitting. When relying on just term and document frequencies we are guided by the following heuristics in the choice of ranking functions: COTs that occur often (per term or document) should be assigned higher weight since a lexicon entry with this COT will be used heavily in feature extraction. On the other hand, we hypothesize that COTs occurring very frequently are likely to have less discriminatory value in sentiment classification. For instance, the COT *Oslo Børs* (*Oslo Stock Exchange*) occurs in the greatest number of documents in our collection, however, it carries no sentiment and, hence, has no discriminatory value as an entry in a sentiment lexicon.

Additionally, we conjecture that if two terms have some meaningful relationship, either semantic or lexical, their tendency to co-occur will be greater than what would happen by pure chance i.e. the distribution is biased. This bias we try to measure with the statistics mutual information and chi-squared and we contemplate that it can be used as an indicator of COT importance in sentiment classification (Matsuo and Ishizuka, 2004). We will now account for the five ranking functions we will employ and evaluate in selecting COTs to annotate.

Term Frequency The simplest statistic we employ in our system is term frequency at the collection level. The COT that occurs the most number of times receives that highest rank using this statistic. This makes sense since annotating COTs that occur more often will be rewarding since that entry in the sentiment lexicon will be used many times in feature extraction. We use this statistic out of its simplicity and its history of yielding satisfactory results in feature selection, even when compared to more complex methods (Manning et al., 2008). Formally, we use the notation

$$tf_{u,v} \tag{1}$$

to refer to the term frequency of the COT composed of term u and v.

Inverse Document Frequency Where the term frequency statistic has value in favouring exhaustivity, the second statistic we employ, the inverse document frequency statistic, gives preference to specificity (Jones, 1972). This seems propitious given the assumption that COTs occurring infrequently across a collection contains much sentiment bearing information, and, hence, is suited for entry in a sentiment lexicon (Robertson, 2004). For instance, the COT *Oslo Børs* (*Oslo Stock Exchange*) occurring in the greatest number of documents in our collection carries no sentiment. Formally, we use the notation

$$idf_{u,v} = \frac{N}{df_{u,v}} \tag{2}$$

to refer to the inverse document frequency of the COT composed of term u and v. N denotes the number of documents in the collection and $df_{u,v}$ denotes the document frequency of the COT composed of term u and v.

TF-IDF A widely used algorithm for indexing in Information Retrieval is tf-idf, initially proposed by Jones (1972). When used as a ranking statistic it weights together the exhaustivity of a term, occurring frequently, with its specificity, occurring infrequently across the collection. Formally, we use the notation

$$tf\text{-}idf_{u,v} = tf_{u,v} \cdot lg\left(idf_{u,v}\right) \tag{3}$$

to refer to the tf-idf statistic of the COT composed of term u and v.

Mutual Information Mutual information measures the mutual dependence of two random variables (Manning et al., 2008) and can, in our case, be used to measure the degree of bias in co-occurrence of two terms (Matsuo and Ishizuka, 2004). Formally, we use the notation

$$mi_{u,v} = \frac{df_{u,v}}{N} \cdot lg\left(\frac{N \cdot df_{u,v}}{df_u\, df_v}\right) \tag{4}$$

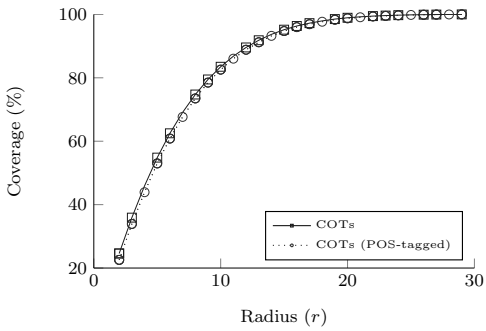
to refer to the mutual information statistic of the COT composed of term u and v.

Chi-Squared In statistics, Chi-squared is used to test the independence of two events (Manning et al., 2008) and can, if the events are the occurrence of terms be used to measure the degree of bias in co-occurrence of two terms, like for mutual information (Matsuo and Ishizuka, 2004). Formally, we use the notation

$$\chi^2_{u,v} = \frac{\left(df_{u,v} - N \cdot df_u \cdot df_v\right)^2}{N \cdot df_u \cdot df_v} \tag{5}$$

to refer to the chi-squared statistic of the COT composed of term u and v.

Fig. 4: Sentence and COT coverage as a function of radius



3.5 Radius (r)

The radius variable input to our system sets, as explained in Definition 3 the permitted distance between two terms within the same sentence to become a candidate COT. An increasing radius will generate more candidate COTs, however, it is likely that the semantic, lexical or other relationship between two terms with great distance between them to be weaker. To make an informed decision on radius range, we computed the coverage (fraction of all COTs in the corpus extracted) for a varying radius. This was done both with and without using a POS-tagger and is shown in Figure 4. This Figure shows that if the radius r is capped at 2 only ∼20% of the COTs in the corpus are extracted. It is furthermore clear that as r approaches 8 the COT (both POS-tagged and unprocessed) coverage moves towards ∼80%. Including a larger radii than this will, we conjecture, yield little benefit, given the increased effort. Hence, we varied the radius between 2 and 8 as inputs to our system (note that the results presented in Table 2 and 3 are reported in increments of 2, due to only slight variation in results between single increments).

3.5.1 Lexicon Size (σ, ρ)

The lexicon size input to our system sets the number of (top-ranked) candidate COTs to be manually annotated in composition of a sentiment lexicon. We varied the lexicon size (σ) both in absolute size, $\sigma \in \{1000, 2000, 4000\}$, and in proportion ($\rho$) to the available candidate COTs in the collection given the radius (larger radii gives more candidate COTs), $\rho \in \{10\%, 30\%, 50\%\}$. Both the absolute and relative measure of lexicon size were included, and used separately, to properly decouple the effect of increasing radius from increasing lexicon size - for small radii the number of candidate COTs is less than 4000. Hence, increasing the radius in this case is, perhaps misleadingly, beneficial since a larger number of candidate COTs are available.

Table 1: Machine Learning Input Features

Feature	Description	Type
PositiveCotsTitle	Number of COTs annotated as positive in sentiment lexicon occurring in title	Discrete
NegativeCotsTitle	Number of COTs annotated as negative in sentiment lexicon occurring in title	Discrete
NeutralCotsTitle	Number of COTs annotated as neutral in sentiment lexicon occurring in title	Discrete
PositiveCotsLead	Number of COTs annotated as positive in sentiment lexicon occurring in lead text	Discrete
NegativeCotsLead	Number of COTs annotated as negative in sentiment lexicon occurring in lead text	Discrete
NeutralCotsLead	Number of COTs annotated as neutral in sentiment lexicon occurring in lead text	Discrete
IsWeekend	Whether article has been published on the weekend	Binary
IsWeekday	Whether article has been published on a weekday	Binary
IsStockExchangeOpen	Whether the article has been published within the opening hours of OsloStockExchange (OSE)*	Binary
IsStockExchangeCategory	Whether article is published in the Stock Exchange category of hegnar.no	Binary
IsAnalysisCategory	Whether article is published in the Analysis category of hegnar.no	Binary
IsEconomyCategory	Whether article is published in the Economics category of hegnar.no	Binary
StockPricePercentage-ChangeYesterday	Percentage stock price change last day of trade	Continuous
StockPricePercentage-ChangeLastWeek	Percentage stock price change last week of trade	Continuous

*Regular trade on the Oslo Stock Exchange (OSE) is Monday through Friday 0900 - 1620 CET

3.5.2 Machine Learning Classifiers (c)

We used three different machine learning classifiers, Naïve Bayes (NB), Random Forest (RF) and J48, in training and classification, to allow for simple comparisons. These particular classifiers were included having been relatively widely used in previous related work on sentiment analysis. Additionally, Random Forest (RF) and J48 are suited for our investigation as both have recently been found to yield the highest performance in this domain (Thelwall et al., 2010; Castillo et al., 2011; Ali et al., 2012). Apart from feature binning, no classifier tuning was performed as this was not of emphasis in this paper.

3.6 Machine Learning Input Features

Given an article document and the sentiment lexicon, acquired based on the input parameters, we counted the number of positive, negative and neutral COTs, which were entries in the lexicon, in the title and lead text of the article. We refrained from using the main text as the basis of the feature vector due to the revealed correlation between title and aggregate sentiment, in addition to between lead and aggregate sentiment classification, to be in the order of $\sim$0.9. Hence, the title and

lead text seem close to adequate in determining the sentiment of the article as a whole.

We additionally added features fixed when varying the input parameters to our system, radius r, lexicon size σ and ranking function f. These relate to when the article in question has been published (weekend, weekday and whether the stock exchange is open), in what category the article has been published (category tag extracted from url of article) in addition to the percentage change in price of the stock that is identified as the main target of the article, both the trading day and last trading week.

These other features were included to give the classification system a state-of-the-art precision level and, hence, giving our approach greater authority. The relative differences from changing the parameter inputs to our system will not be affected by these additional fixed features. A complete list of the features used in our system with description and data type is presented in Table 1.

4 Experiments and Evaluations

The domain of our interest, as mentioned, is the news flow of stocks listed on the Oslo Stock Exchange (OBX) and, hence, we selected the online financial news publisher hegnar.no as the source of our dataset. We extracted 992 articles from their website, all news articles were published in the time period from Nov 29^{th} 2012 to Jan 15^{th} 2013 and with exactly one ticker (symbol consisting of three or more letters uniquely identifying a stock) included in the article tag set, and annotated the aggregate sentiment of each of these articles being directed towards the stock in question, as either positive, negative or neutral/objective. Articles with no tickers are assumed not to direct sentiment towards any particular stock. Articles with more than one ticker, requires reasoning about which parts of the article are directed towards each of the tagged stocks, well outside the scope of this paper. These annotated documents were then POS-tagged in input to our system as *Annotated article data set D*, in Figure 1.

We then ran our system with all possible combinations of the input parameters radius $r = \{2, 4, 6, 8\}$, ranking function $f = \{tf, idf, tfidf, mi, \chi^2\}$, machine learning classifier $c = \{nb, rf, j48\}$ and lexicon size, either specified absolutely $\sigma = \{1000, 2000, 4000\}$ or relative to the number of candidate COTs $\rho = \{10\%, 30\%, 50\%\}$. This required us to manually annotate the top-ranked list of candidate COTs for each combination of radius, ranking function and lexica size. In total, 7990 COTs were manually annotated, revealing that there is considerable overlap between the different candidate COT list rankings. The effort required to manually annotate this combined lexica was ∼4 hours - negligible compared to the time spent manually annotating the 992 articles composing the test and training data.

The results from running the system with the different input parameter combinations for the absolutely specified lexicon size ($r \times \sigma \times f \times c$) and the relative lexicon size ($r \times \rho \times f \times c$) are given in Tables 2 and 3, respectively. Graphical visualizations of the results are presented in Figures 5a through 6i.

The overall precision of the system, as evident in Tables 2 and 3, approaches approaches ∼70% - the highest achieved precision was 69.1%, highlighted in bold in Table 3. This is comparable to the state-of-the-art classification precision and

Table 2: Classification Precision Results With Absolute Lexicon Size

		$\sigma = 1000$				$\sigma = 2000$				$\sigma = 4000$			
MLC	**Ranking**	**COT radius =**				**COT radius =**				**COT radius =**			
	Function	**2**	**4**	**6**	**8**	**2**	**4**	**6**	**8**	**2**	**4**	**6**	**8**
NB	*tf*	60.1	50.9	52.9	52.8	64.3	60.4	58.1	57.3	63.9	65.1	63.7	60.3
	idf	61.0	59.0	58.8	57.0	63.6	62.2	60.7	60.3	64.5	65.4	64.1	63.1
	tfidf	59.2	55.0	53.6	51.2	63.4	60.8	59.1	59.3	64.4	65.1	62.2	60.9
	mi	60.5	58.0	55.4	55.9	64.5	62.0	61.5	61.0	64.9	65.0	64.2	62.1
	χ^2	57.5	50.0	50.2	49.0	63.5	53.1	53.4	54.9	64.0	64.2	60.9	57.9
RF	*tf*	61.7	59.1	59.0	59.4	64.6	62.9	61.5	60.8	65.0	65.8	64.3	63.3
	idf	61.1	59.1	58.5	56.9	63.6	62.4	60.8	60.1	64.8	65.6	64.6	63.3
	tfidf	61.7	60.1	59.9	58.8	64.6	62.2	61.6	61.8	64.9	65.6	64.3	63.3
	mi	61.7	60.7	60.2	60.0	64.8	63.2	62.8	62.1	64.5	65.7	64.7	63.4
	χ^2	59.7	57.5	56.5	55.6	64.8	60.4	57.8	57.9	64.6	65.8	63.5	61.8
J48	*tf*	61.9	59.8	59.4	60.1	64.6	63.1	61.5	61.2	64.9	65.5	64.3	63.6
	idf	61.0	59.2	58.6	57.1	63.8	62.3	60.6	60.3	64.9	65.6	64.6	63.3
	tfidf	61.9	60.3	60.0	59.5	64.5	63.1	61.8	61.8	64.9	65.7	64.3	63.7
	mi	61.7	60.8	60.4	60.2	64.8	63.0	62.8	62.2	64.9	**65.9**	64.7	63.6
	χ^2	60.2	57.8	56.7	56.4	64.8	61.0	58.4	57.3	64.9	65.7	63.4	61.6

close to the human baseline in the domain of sentiment analysis, given the subjective nature of the classification task. Furthermore, tuning of the machine learning classifiers, which was not of the emphasis of this paper, yielded a maximum result of 72.5%. This precision is in line with the state-of-the-art and above results the most comparable related work by Mourad and Darwish (2013), as discussed in Section 2, at 71.1%. We maintain that our task is inherently more difficult than that of the mentioned authors since it performs 1) ternary as oppose to binary classification and 2) document-level in contrast to sentence-level analysis. This affirms the validity of our approach of attaining a satisfactory machine learning classifier in a completely new domain in absence of lexical resources, such as a sentiment lexicon, and restricted dataset size. The manual lexicon acquisition, deemed infeasible by Feldman (2013) was arguably constructed with permissible levels of effort.

Considering the different input parameters the effect of increasing the radius seems dubious at first sight. When lexicon size is absolute, the effect of increasing this variable is negative on classification precision, as evident in Table 2 and Figures 5a to 5i, whereas when the lexicon size is relative to number COT candidates the effect seems positive, clear in Table 3 and Figures 6a to 6i. When performing sentiment analysis with our approach, a goal should be to keep the number of sentiment lexicon entries to annotate at a minimum. Hence, the case lexicon size is absolute should be attached the most weight and the radius should be kept low for the highest classification precision performance for the least amount of manual sentiment lexicon annotation. Intuitively, this finding aligns with the conjecture that the semantic, lexical or other relationship between terms are stronger when they are closer together.

When it comes to lexicon size, σ in absolute and ρ in relative terms, it is clear, not all surprising, that increasing the size yields higher classification precision. Interestingly, the precision enhancement from increasing the size is quickly diminishing - average precision increase from doubling the size from $\sigma = 1000$ to $\sigma = 2000$ is $\sim$6% whereas a doubling from $\sigma = 2000$ to $\sigma = 4000$ is merely $\sim$4%.

Table 3: Classification Precision Results With Relative Lexicon Size

		$\rho = 10\%$				$\rho = 30\%$				$\rho = 50\%$			
MLC	**Ranking**	**COT radius =**				**COT radius =**				**COT radius =**			
	Function	**2**	**4**	**6**	**8**	**2**	**4**	**6**	**8**	**2**	**4**	**6**	**8**
NB	*tf*	63.9	55.7	55.5	58.0	63.6	65.4	65.3	65.7	64.8	68.3	66.6	65.2
	idf	62.8	62.8	62.8	62.8	63.8	62.8	65.1	64.8	65.9	64.2	65.6	67.4
	tfidf	63.0	57.8	56.3	58.7	62.6	65.3	65.1	66.7	64.3	68.1	66.7	67.3
	mi	59.3	61.6	59.1	60.5	61.8	66.0	65.6	65.4	66.0	68.5	68.9	68.9
	χ^2	62.8	60.0	60.7	62.8	59.6	61.4	60.8	60.4	59.4	57.5	60.9	65.4
RF	*tf*	64.7	62.4	62.9	63.7	64.0	65.5	66.6	65.6	65.7	69.0	66.6	67.0
	idf	62.8	62.8	62.8	62.8	63.8	62.8	65.0	65.3	66.2	64.9	66.4	66.9
	tfidf	64.4	63.2	62.9	63.6	64.2	66.4	65.0	66.7	65.4	68.9	66.4	67.0
	mi	62.8	63.3	64.5	65.2	62.7	65.6	66.4	65.4	65.2	68.7	68.9	68.7
	χ^2	62.8	62.8	62.8	62.8	62.8	62.1	62.6	63.1	62.7	64.6	63.9	65.1
J48	*tf*	65.1	62.7	62.7	63.7	63.8	65.9	66.3	65.6	65.9	68.9	66.8	67.6
	idf	62.8	62.8	62.8	62.8	63.8	62.7	64.9	65.3	65.5	64.8	66.6	67.4
	tfidf	64.4	62.9	62.9	63.3	63.9	66.5	64.7	66.9	66.2	69.0	66.5	67.6
	mi	62.7	63.4	64.7	65.2	62.7	65.6	65.9	65.9	65.4	68.7	**69.1**	68.9
	χ^2	62.8	62.8	62.8	62.8	62.8	62.8	62.6	63.0	62.6	64.8	63.9	65.0

Out of the ranking functions *f mutual information* yielded the highest precision - exhibiting ∼4 higher average precision than the lowest performing ranking statistic. We theorize that this is due to this statistic's ability to measure and reward distribution bias in COTs. Lastly, comparing machine learning classifiers (c) it is clear that J48 resulted in the highest precisions, closely followed by Random Forrest (RF). The former resulted in a ∼3% higher precision than the poorest performing Naïve Bayes (NB). This is precisely in line with some of the more recent work on machine learning classifier comparisons within sentiment analysis (Thelwall et al., 2010; Castillo et al., 2011; Ali et al., 2012).

5 Conclusions And Further Work

In this paper we have 1) impugned the supposition that manual lexicon acquisition for sentiment classification is infeasible given the required annotation effort by presenting a novel system attaining state-of-the-art precision with permissible levels of effort and 2) optimized supervised sentiment lexicon acquisition through experimentation and end-task evaluation. Our proof-of-concept system achieved classification precisions up to 72.5%, exceeding the closest comparable work of Mourad and Darwish (2013) obtaining a precision of 71.1% using the ArabSenti lexicon on tweets, while demanding significantly lower annotation efforts than peer lexica (∼4 hours compared to ∼27 hours by the same authors). Hence, our system affirms that one can, with fairly low annotation efforts, achieve satisfactory document-level sentiment classifier, exhibiting state-of-the-art precision, when presented with a new corpus in domain with no pre-existing suited sentiment lexicon - an approach previously having been deemed inefficacious (Feldman, 2013).

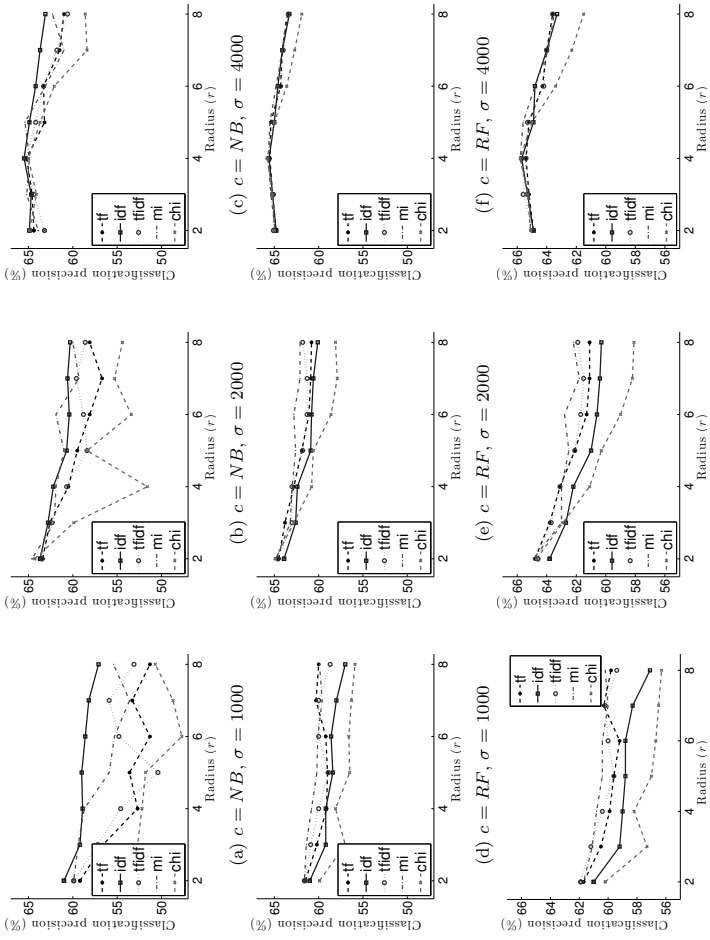
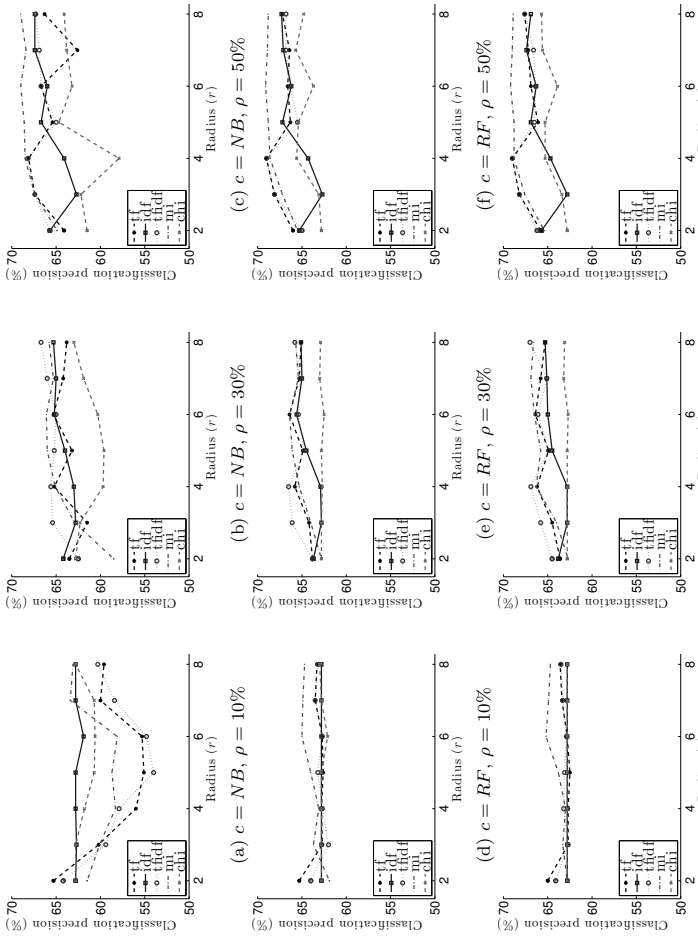


Fig. 5: Classification precision with varying radius, ranking function, absolute lexicon size and machine learning classifier



(a) $c = NB$, $\rho = 10\%$ (b) $c = NB$, $\rho = 30\%$ (c) $c = NB$, $\rho = 50\%$

(d) $c = RF$, $\rho = 10\%$ (e) $c = RF$, $\rho = 30\%$ (f) $c = RF$, $\rho = 50\%$

(g) $c = J48$, $\rho = 10\%$ (h) $c = J48$, $\rho = 30\%$ (i) $c = J48$, $\rho = 50\%$

Fig. 6: Classification precision with varying radius, ranking function, relative lexicon size and machine learning classifier

Additionally, we have optimized supervised sentiment lexicon acquisition by experimenting with different methods of selecting COTs to annotate for sentiment analysis of financial news. In doing this we have used our complete system for conducting supervised sentiment analysis on an annotated dataset in a new domain and answered the questions of how many COTs (absolute and in proportion to corpus size) need to be annotated in order to achieve satisfactory sentiment classification precision. Moreover, we have experimented with which radius (permitted number of words between the COTs) should be used in addition to how the COTs candidate for annotation and entry in a sentiment lexicon should be ranked. In our analysis we find the ranking function *mutual information* to perform the best while precision increases diminishingly with lexica size and weakens with larger radii.

Although we argue our presented approach requires permissible levels of manual annotation effort, we will still be investigating ways of minimizing this effort by iteratively using the annotated sentiment lexicon to select new COTs for entry in the same lexicon going forward. As an additional venue for further work, we will be evaluating the pragmatic value of our developed system by assessing whether it can make statistically significant, and remunerative, predictions of stock price development on the Oslo Stock Exchange.

Acknowledgements

This work is partially funded by NxtMedia (www.nxtmedia.no) and the Telenor Group (www.telenor.com).

References

Muhammad Abdul-Mageed, Mona T Diab, and Mohammed Korayem. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 587–591. Association for Computational Linguistics, 2011.

Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 2012.

Fahad Alotaiby, Ibrahim Alkharashi, and Salah Foda. Processing large arabic text corpora: Preliminary analysis and results. In *Proceedings of the second international conference on Arabic language resources and tools*, pages 78–82, 2009.

Sarah L Andrews, Peenaki Dam, Damien Frennet, Summit Chaudhuri, Ricardo Rodriguez, Ashok Ganapam, Frank Schilder, and Jochen Lother Leldner. Methods and systems for generating composite index using social media sourced data and sentiment analysis, December 27 2011. US Patent App. 13/337,662.

Michelle Annett and Grzegorz Kondrak. A comparison of sentiment analysis techniques: Polarizing movie blogs. In *Advances in artificial intelligence*, pages 25–35. Springer, 2008.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

Alexandra Balahur, Ralf Steinberger, Mijail Alexandrov Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. In *LREC*, 2010.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*, 2013.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*, 2007.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233, 2006.

Simon Clematide and Manfred Klenner. Evaluation and extension of a polarity lexicon for german. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, page 7, 2010.

Sanjiv R Das and Mike Y Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

Bonnie J Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. Improved word-level alignment: Injecting knowledge about mt divergences. Technical report, DTIC Document, 2002.

Eduard C Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1761–1764. ACM, 2010.

Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

Johannes Fürnkranz. A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, 3(1998):1–10, 1998.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7, 2007.

Benjamin Graham. *The Intelligent Investor, A Book of Practical Counsel*. Harper & Brothers Publishers, 1959.

Benjamin Graham, David Le Fevre Dodd, and Sidney Cottle. *Security analysis*. McGraw-Hill New York, 1934.

Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Bernardo A Huberman and Lada A Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

J Kamps, MJ Marx, RJ Mokken, and M de Rijke. Using wordnet to measure semantic orientations of adjectives. In *Proceedings of fourth international conference on Language Resources and Evaluation,{LREC}, Vol IV,*, pages 1115–1118. European Language Resources Association (ELRA), 2004.

Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics, 2006.

Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Ahmed Mourad and Kareem Darwish. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. *WASSA 2013*, page 55, 2013.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Sentiful: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, Wei Wei, and Jon Atle Gulla. Evaluating feature sets and classiers for sentiment analysis of financial news. To appear in the proceedings of the 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2014). Available at: `http://folk.ntnu.no/palchrnj/papers/paper_wic2014.pdf`, 2014.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

Mihir Parikh, Robert M Fabricant, and Ed Hicks. Sentiment analysis, January 19 2012. US Patent App. 13/353,982.

Wei Peng and Dae Hoon Park. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. *Urbana*, 51:61801, 2004.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. Learning sentiment lexicons in spanish. In *LREC*, pages 3077–3081, 2012.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In *IJCAI*, volume 9, pages 1199–1204, 2009.

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

Victoria L Rubin. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144. Association for Computational Linguistics, 2007.

Diana Santos. Punctuation and multilinguality: Some reflections from a language engineering perspective. *Working Papers in Applied Linguistics*, 4(98):138–160, 1998.

Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL*, pages 300–307, 2007.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. *The General Inquirer: A Computer Approach to Content Analysis.* MIT press, 1966.

Songbo Tan, Yuefen Wang, and Xueqi Cheng. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 743–744. ACM, 2008.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics, 2012.

Wei Wei and Jon Atle Gulla. Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 404–413. Association for Computational Linguistics, 2010.

Wei Wei, Ole J. Mengshoel, and Jon Atle Gulla. Stochastic search for effective features for sentiment classification. 2013.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM, 2005.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3): 165–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE, 2003.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48, 2008.

(This page is intentionally left blank.)

Chapter 8

Paper III

Pål-Christian Salvesen Njølstad, Lars Smørås Høysæter, Øyvind O. Salvesen, and Jon Atle Gulla: *The interaction impact of firm-specific news and market wide sentiment on stock price behavior - evidence from the Oslo Stock Exchange (OSE)*, working paper planned submitted to a relevant journal in empirical or quantitative finance.

(This page is intentionally left blank.)

The interaction impact of firm-specific news and market-wide sentiment on stock price behavior - evidence from the Oslo Stock Exchange (OSE)

Pål-Christian S. Njølstad, Lars S. Høysæther,
Øyvind O. Salvesen, and Jon Atle Gulla
Norwegian University of Science and Technology

ABSTRACT

We use Autoregressive conditional heteroskedasticity (ARCH) models to link firm-specific news sentiments and an aggregate market-wide sentiment index to the return, volume, and order size of ten stocks listed on the Oslo Stock Exchange (OSE). We find that positive articles predominantly lead to significant increases in volume while negative articles have the opposite effect. The same appears to be the general proclivity for order size. Only negative articles are found to significantly impact return, leading to a reduced subsequent stock price. The interaction between news articles and market-wide sentiment is also found to be statistically significant. Although the sign of this effect seems firm-idiosyncratic, our analysis reveals that white chips' reactions are of greater magnitude than that of blue chips. Albeit evidence is found for significant interaction impact on return, our low-resolution daily data, supposedly subject to reverse causality caused endogeneity, can only be used to draw tentative conclusion on this, advocating further analysis with high-frequency data. Lastly, the sentiments of these firm-specific news were determined using a specially devised sentiment engine. As such, this paper ultimately serves as a strong validation of this system.

JEL classification: G12, G14, C32.

Financial support was received from NxtMedia and the Telenor Group to carry out the thesis that ultimately culminated in this paper. The authors are grateful technical help from Jon Espen Ingvaldsen and Arne Dag Fidjestøl. All errors remain our own.

The historical vicissitude of the stock market remains a puzzle from the viewpoint of a rational investor, if there truly ever existed one. The markets have exhibited a pattern that relentlessly alternates between bubbles and recessions, even though the intrinsic values[1] of the underlying assets being traded remains much more stable, as discussed by Shiller (2000). For instance, the same author compares historical valuations and earnings of the S&P Composite Stock Price Index in the time period 1871 to 2000, and it is clear that the former, which is a function of both the irrational investor's exuberance as well as lethargy, fluctuates far more than the latter.

Furthermore, there is amble evidence of the financial markets both over and under-reacting significantly to non-events (Fama, 1998). A recent example of such overreaction is from the Oslo Stock Exchange (OSE), the subject of this paper, that occurred the 14^{th} and 15^{th} of October 2008. On this first day, the Oslo Stock Exchange All Shares Index (OSEAX) first rose 11.6% from 304.4 to 339.8 points before subsequently falling 5.8% to a close of 320.2. The following day, the index fell another 8.2% to 293.9, the by far greatest fall of all the European stock exchanges. This equates to values worth NOKb ∼110 having been created and subsequently NOKb ∼140 being destructed - in just over 30 hours. Although the stock market was supplied with some new information these two days, including Ben Bernanke, then-Chair of the Federal Reserve, trying to restore confidence in the credit market, a slight adjustment of the Norwegian InterBank Offered Rate (NIBOR), and publication of disappointing U.S. consumer confidence numbers, this sudden swift shift from ebullience to discouragement by far defies that what could be rationally expected. As the Norwegian News Agency (www.ntb.no) put it upon OSE close, the fall was "(...) in large caused by fear of the severity of the global economic recession (...)". This fear, that the news agency makes reference to, is commonly, and in more general terms, referred to as *investor sentiment*. This can broadly be defined as the beliefs about future cash flows and investment risks that cannot be factually justified (Baker and Wurgler, 2007). This is, by many researchers, deemed the cause of these seemingly unexplainable, irrational market movements.

However, one could argue that these erroneous beliefs of irrational investors, causing deviations from intrinsic values, simply should represent an opportunity for arbitrageurs to make a sound, risk-free profit. This is what the standard theoretical models, adjusted for imperfect information and transaction costs, would predict for any such mispricing. In fact, the arbitrageurs should, in conformance with the Efficient Market Hypothesis (EMH), be able to exactly offset whatever effect these unjustifiable sentiments should have on asset valuations (Pontiff, 2006). On the contrary, as Shleifer and Vishny (1997) emphasizes, betting against these sentimental investors has turned out to be both costly and risky and, hence,

[1] Value determined through fundamental analysis and, often, decoupled from current market value.

arbitrageurs have not been able to drive prices back to fundamentals, as predicted by the named models (Baker and Wurgler, 2007). There is great uncertainty associated with when irrational mispricings will revert to fundamentals, and this is a problem since, as John Maynard Keynes famously put it, "markets can remain irrational longer than you can remain solvent" (Lowenstein, 2000). Furthermore, as Shleifer and Summers (1990) present ample evidence of, these "noisy traders" not only contribute to mispricing but are able to earn higher than expected returns. If at first sight surprising, this makes perfect sense since these investors any other way would be supplanted by the arbitrageurs and disappear from the market.

If the rational investor needs to abide by this irrational investor sentiment, the question that then remains is how such an investor best can, if not by emulating the behavior of a "noisy trader", systematically exploit the evident relationship between investor sentiment and stock price valuations. This question has been the subject of much research in recent times. In order to exploit this relationship, one necessarily must know how to quantitatively measure investor sentiment and its effects. The predominant part of the literature uses either 1) sentiment proxies, 2) macroeconomic news, and / or 3) company-specific earnings announcements for this measurement (Groß-Klußmann and Hautsch, 2011). As Baker and Wurgler (2007) note, an exogenous shock in investor sentiment can lead to a chain of events further enforcing the market investor sentiment in a positive feedback loop. For example, if the current Chair of the Federal Reserve, Janet Yellen, through one of her speeches succeeds in inducing positive investor sentiment, this might instantaneously cause increased trading activity. If this drives up share prices, this will lead to more investor confidence, and yet further enhanced investor sentiment. As such, and as the aforementioned authors take heed of, this sentiment could in principle be observed at any, or every, part of this positive feedback chain - either through low-level linguistic analysis of Yellen's speech or through the instantaneously resulting increased market liquidity. However, not all parts of this positive feedback chain are equally well suited as a proxy of investor sentiment, for multiple reasons. First of all, since not all of these proxies are easily observable to investors their effect are likely to be limited - it is ultimately the sentiment of *most* investors (as in with the *most* ability to determine asset prices) that we wish to understand. Secondly, measuring investor sentiment further from this exogenous shock will likely make the possibility to profit from the change in investor sentiment to be forgone.

In this paper, we argue that financial news is a most felicitous source for measuring investor sentiment. Firstly, this is, in effect, a funnel of investor sentiments - journalists and editors of financial news aggregates sentiments from a multitude of other sources, amplifying the prominent and filter out the irrelevant, resulting in an easily processed, widely-read

source of information for market agents. Secondly, if employing sophisticated linguistic machine-aided analyses one can act on sentiments expressed in the news before the market reacts and, hopefully, achieve superior returns in this manner. And thirdly, if using news as a source of sentiments one has the ability to perform analysis of firm-specific sentiments. Since it is much more likely that a single stock is mispriced than an entire index, this is again deemed more likely to lead to remunerative rewards.

In order to analyze financial news, and its likely effects on investor sentiments and stock price reactions, we have developed a sentiment engine combining computational linguistics and machine learning. It classifies the sentiments expressed in financial news articles directed towards firms whose equity is traded on the OSE. This system has a precision of $\sim$70%, which is in line with the state-of-the-art in this subjective, imprecise task of sentiment classification (Balahur et al., 2010; Mourad and Darwish, 2013). The reason for narrowing our analysis to the news flow associated with stocks listed on the OSE is twofold: 1) as Gjerde and Saettem (1999) establish, the Norwegian financial market is less mature compared to that of the U.S. and others, presumably leaving greater opportunities for exploiting sentiment reactions, and 2) the linguistic barriers[2] to building such a sentiment engine for the Norwegian language are significant, presumably leaving a greater upside if successfully able to do so.

We have used our developed sentiment engine to classify the news flow of ten companies listed on the OSE over a period of six years (from the 4^{th} of January 2008 to 14^{th} of April 2014). These have been selected based on market value and liquidity such that five of them are so-called blue chips (stocks with high market value and liquidity) and the remaining five are, what we will refer to as white chips (the least valuable poker chip - stocks with low market value and liquidity). These will be described in detail in section II.

We use the output of this sentiment engine, the number of articles with a positive, neutral, and negative sentiment classification published per day, week and month, to predict 1) return, 2) volume (traded monetary value) and 3) average order size on a daily, weekly and monthly basis for the ten companies being studied. To rule out other explanatory factors we include a list of control variables in our regression models. We study whether the reactions to the sentiments expressed in news differs significantly between the two different types of firms considered and, ultimately, if it differs between specific firms. For our predictions we use the Autoregressive conditional heteroskedasticity (ARCH) model, well-suited for for financial time-series.

Furthermore, we study the interaction impact of firm-specific news and aggregate market-wide sentiment ('Bull' or 'Bear' market sentiment) on stock price behavior. With this, we

[2]The availability of lexical and linguistic resources is much more limited for Norwegian than for the English language.

hypothesize that investors react differently in bullish and bearish times, and that this is vital to be aware of when trying to systematically exploit the relationship between investor sentiment and stock price development. Additionally, and as a sanity check to our developed system, we investigate whether news published during holidays (when the OSE is open) lead to different stock behavior reactions than news published at other times. This seems probable since investors are less likely to follow the news closely and to be trading during the holiday season.

The main contributions of this paper are as follows: 1) positive articles predominantly lead to significant increases in volume while negative articles have the opposite effect, 2) positive articles predominantly lead to increased order sizes, while the converse is true for negative articles, 3) negative articles are found to significantly impact return, which will lead to reduced subsequent returns, 4) the interaction between news articles and the aggregate marked-wide sentiment is also found to be statistically significant and, although the sign of this effect seems firm-idiosyncratic, white chips' reactions are of greater magnitude than that of blue chips, and, 5) the significance of our regression models using daily data, despite supposedly suffering from the endogeniety problem, reveal strong relationships between intraday return and news flow. This justifies more granular investigations with high-frequency data. Additionally, and perhaps in more excitement to the computer scientist than the financier, this paper ultimately serves as a strong validation of our in-house, specially developed sentiment engine.

The remainder of this paper is organized as follows. In the next section, we account for previously conducted work related to ours. We then present the data used in this study along with descriptive statistics in Section II. In the subsequent section, III, we discourse the methodology used. In Section IV we present and discuss our results and, finally, in Section V we provide concluding remarks and suggest compelling venues for further work.

I. Related Work

As mentioned in the previous section, the relationship between investor sentiments and stock price behavior has been subject to much research, and the predominant part of the literature uses either 1) sentiment proxies, 2) macroeconomic news and / or 3) company-specific earnings announcements for measuring these sentiments. These are then linked to stock returns and / or different liquidity measures, hoping to be able to make predictions about future behavior (Groß-Klußmann and Hautsch, 2011).

Examples of sentiment proxies include, as nicely reviewed by Baker and Wurgler (2007), results in soccer and other sports games (Edmans et al., 2007), investor surveys (Brown

and Cliff, 2004), investor moods (Kamstra et al., 2003; Kim and Park, 1994), retail investor trades (Barber et al., 2006), mutual fund flows (Frazzini and Lamont, 2008; Brown et al., 2003), trading volume (Scheinkman and Xiong, 2003), dividend premia (Baker and Wurgler, 2004), close-end fund discount (Neal and Wheatley, 1998), option implied volatility, through measurement of the VIX[3] (Whaley, 2000), IPO frequency and volume (Ljungqvist et al., 2006), equity issues over total new issues (Baker and Wurgler, 2000), and levels of insider trading (Seyhun, 2000). More recently, the Harvard MBA Indicator[4], developed by the former banking analyst Ray Soifer, has gained some popular media attention and is currently predicting that stock prices will be depreciating. Since we are studying the interaction between firm-specific news and the aggregate market-wide sentiment, we will be using a handful of these proxies to devise a simple sentiment-index for stocks traded on the OSE.

Numerous studies have also sought to link company announcements and link macroeconomic news to stock price behavior. The former was done by Malatesta and Thompson (1985) who studied the impact of (partially anticipated) corporate acquisitions on subsequent stock prices. Landsman and Maydew (2002) investigated whether quarterly earnings announcements cause abnormal changes in a stock's volume and return volatility. A similar study was done for (anticipated and unanticipated) dividend announcements (Graham et al., 2006). The impact of macroeconomic news on interest rates (Ederington and Lee, 1993), foreign exchange rates (DeGennaro and Shrieves, 1997), and government bonds (Fleming and Remolona, 1999; Hautsch and Hess, 2002) have also been studied.

Most previously published work on investor sentiment, especially those using aggregated sentiment proxies, limit their scope to weekly and / or monthly time horizons. Few studies link changes in sentiment to intraday stock price behavior. Berry and Howe (1994) use the number of news releases by the Reuter's News Service to forecast intraday market activity and find a positive, moderate relationship between this public information and trading activity. A related study has affirmingly been done using the news announcements reported daily by Dow Jones & Company (publishing and financial information firm) (Mitchell and Mulherin, 1994).

The two most elaborate investigations on investor sentiment and intraday activity have been done by Kalev et al. (2004) and, more recently, Groß-Klußmann and Hautsch (2011). The former relates firm-specific announcements to volatility using high-frequency data from the Australian Stock Exchange. The authors use ARCH-type regression models for predicting

[3]The Chicago Board Options Exchange Market Volatility Index, or sometimes known as the "investor fear of gauge", measuring the implied volatility of S&P 500 index options (Whaley, 2000).

[4]The indicator simply recommends buying stocks as long as less than 10% of Harvard MBA graduates take "market sensitive positions" (e.g. investment banking, venture capital, equity sales and trading etc.) and to sell when this figure is above 30%.

volatility - the same methodological approach as we take in this paper. The latter, which is the work most closely related to ours, relates articles published by Reuters, and tagged with a positive, neutral, or negative sentiment by the Reuters NewsScope Sentiment Engine, with stock return, volume, volatility, depth, and bid-ask spread. In doing so the authors use high-frequency data on stocks listen on the London Stock Exchange (LSE). The authors find strong volume and volatility reactions, and that they are widely stable across the market. Less distinctive reactions are found for depth and bid-ask spread. Evidence of abnormal high returns after published news sentiments are also revealed.

Groß-Klußmann and Hautsch (2011) limited their study to liquid stocks only. Illiquid stocks are, as some will argue, of paramount interest in such an analysis since they more often than liquid stocks tend to be mispriced (Lee and Swaminathan, 2000). Hence, we take special care to handle blue chip and white chip stocks (as the two types of stocks are referred to as in this paper) separately. Furthermore, we study the interaction impact of firm-specific news and the aggregate market-wide sentiment on stock return, volume, and order size. The impact interaction between aggregate market-wide sentiment and earnings announcements was examined by Mian and Sankaraguruswamy (2008), and the impact of the interaction between aggregate market-wide sentiment and a surprise corporate takeover bid on bidder stock price reaction has also been documented (Rosen, 2006). However, no work has, to our knowledge, compared the general market mood and firm-specific news with regards to stock price behavior. Additionally, the impact of the interaction between firm-specific news and seasonality (holiday versus no holiday) on the same dependent variables, which we investigate, is thus far undocumented.

In coda, our work is novel in three ways. Firstly, we evaluate our in-house developed sentiment engine and prove its prowess in predicting stock price behavior. Secondly, we compare and contrast the impact firm-specific news has on blue chip and white chip type stocks. And thirdly, we study the impact of interactions between aggregate market-wide sentiment, firm-specific news, and seasonality on stock prices.

II. Data

As the number of published opinionated, relevant financial articles grows, it becomes practically impossible to manually monitor these in an effective manner. Keeping up with the news flow of a publicly listed stock is of interest for any of its stakeholders - be it an active investor, portfolio manager, financial analyst, bond holder, tax authority or competitor. Hence, research addressing the scalability of this problem has emerged, and soon evolved, drawing on theory from the intersection of machine learning, computational linguistics (Wei

and Gulla, 2010) and, of course, financial economics. These days numerous software tools for helping its usesr automatically monitor the sentiments from this panoply of sources are readily available[5]. Such tools analyze the information conveyed in Internet documents using linguistic pattern recognition algorithms and are, often, coupled with machine learning techniques. Since this paper is focused on stocks listed on the OSE and the associated news flow of its stocks are written in Norwegian no ready-made software tools are easily available for our purposes, since most, if not all, of these are tailored for the English language[6]. Hence, we have developed a sentiment engine customized for our purposes. This engine is briefly described in Section II.B.

Input to this engine are 9,476 articles from the online version of the financial news Finansavisen, Hegnar.no (www.hegnar.no). This is in effect all news published between the 4^{th} of January 2008 and the 14^{th} of April 2014 on the ten selected stocks. The details of these selected stocks and their associated news flows are given in Sections II.A.1 and II.A.2, respectively. We also input a set of control variables (listed and described in Section II.C) to our analysis and create a sentiment index (explained for in Section II.C) for the study of interaction effects.

A. Sources

For our analysis we need information on sentiments, stock price behavior reaction in addition to information on other potential explanatory factors, so as to avoid the omitted variable bias problem, common in regression analysis. The sources of this information will now be accounted for in turn.

A.1. Stock Price Behavior

As previously noted, we selected ten stocks listed on the OSE for our analysis. These were the top five and bottom five companies by market value meeting the minimum requirement of an associated news flow of at least 200 articles published on Hegnar.no in the six-year period considered. In addition to studying stock price returns, it is of interest to examine the volume and average order size of these stocks. The former to learn about the extent to which it is actually possible to carry out trades predicted by a regression model, and the latter to understand which investors react to information contained in the news. Order size

[5]FinSentS (www.finsents.com), SNTMNT (www.sntmnt.com), and OPFINE (www.jane16.com) are some of these tools (that, incidentally, are highly ranked by Google search).

[6]Although attempts have been made to develop language-agnostic sentiment engines, like that of Evans et al. (2007), progress in this sub-field of sentiment analysis remains very limited.

is likely to go down if retail investors react strong relative to institutional investors, and vice versa. Some descriptive statistics of these ten selected stocks is given in Table I.

Ticker	Market value [NOKb] (% of OSE)	Average intraday return [%]	Average daily traded monetary value [$NOKm$]	Average daily order size [NOK]
FUNCOM	0.4 (0.02)	0.051	5.6	15869
IOX	0.3 (0.02)	-0.047	1.7	20619
NAUR	0.9 (0.01)	-0.169	0.8	1207
NOR	0.2 (0.06)	-0.191	10.7	38350
NSG	1.0 (0.05)	-0.033	24.9	24688
RCL	65.5 (3.53)	0.079	129.1	77508
SDRL	92.4 (4.99)	0.069	363.4	85424
STL	539.7 (29.13)	0.020	1130.6	173060
TEL	195.5 (10.55)	0.025	352.4	89049
YAR	72.8 (3.93)	0.048	627.8	91131

Table I Market value, average intraday return, average daily traded monetary value and average daily order size statistics by ticker

A.2. News Flow

The articles extracted from Hegnar.no, on the ten stocks detailed in the previous section, were input to our sentiment engine and classified as carrying either positive, neutral or negative sentiment towards the ticker (or tickers) in their tag set. The historical total number of published articles with at least one of the ten selected stocks in its tag set is shown in Figure 1.

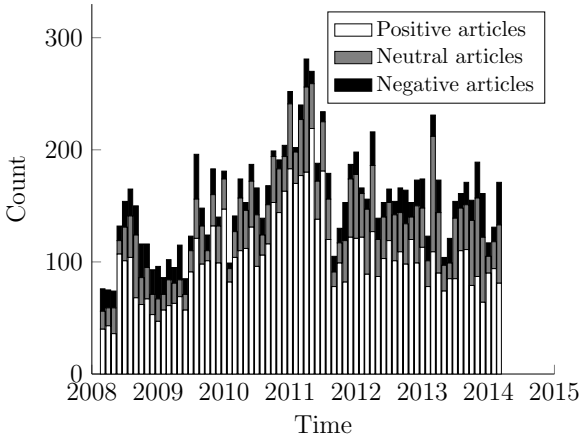


Figure 1. Historical publication count for all tickers considered

Perhaps not very surprising, the news flow associated with the ten stocks varies greatly, from the blue chip, national oil company Statoil (STL) to the white chip, speculative video game

developer Funcom (FUNCOM). The publication count for the period, per ticker, is given in Table II. This is further broken down into the number of articles with positive, neutral, and negative sentiment in the same table.

Ticker	Total number of articles published	Positive articles published (% of total)	Neutral articles published (% of total)	Negative articles published (% of total)
FUNCOM	330	115 (35)	92 (27)	123 (38)
IOX	223	102 (46)	47 (21)	74 (33)
NAUR	339	148 (44)	80 (24)	111 (32)
NOR	441	189 (43)	130 (29)	122 (28)
NSG	656	234 (36)	178 (27)	244 (37)
RCL	627	307 (49)	188 (30)	132 (21)
SDRL	1483	571 (39)	540 (36)	372 (25)
STL	4228	1394 (33)	1612 (38)	1222 (29)
TEL	1510	574 (37)	536 (36)	400 (27)
YAR	1570	595 (38)	555 (36)	420 (26)

Table II Total number of published articles, and the classification of these, by ticker

A.3. Aggregate Market-Wide Data

To correctly control for other likely determinants of stock price behavior, we gathered data on OSEAX historical return and volatility, interest rates, commodity prices, exchange rates, and financial reporting. These were collected from the financial databases Ecowin and Bloomberg, the websites of Norges Bank (the central bank of Norway: www.norges-bank.no), and from the OSE's official reporting portal NewsWeb (www.newsweb.no). OSEAX historical return and volatility data was included as the levels of these variables can be reasonably expected to effect investor reaction to news sentiments. The same goes for interest rate, the three-month Norwegian InterBank Offered Rate (NIBOR), which reflects the current risk-free return in the market. The Brent Crude oil price acts as a natural measure for the current state of the highly oil-dependent Norwegian economy. The EUR-NOK exchange rate has been included since the European Union is, by far, Norway's most important trade partner (judging by both import and export numbers published by Statistics Norway: www.ssb.no). Lastly, we included financial reporting announcements submitted via NewsWeb, to control for the well-documented effect this has on stock price behavior (Kross and Schroeder, 1984). The historical development of these variables, over the six-year period in question, are all visualized in Appendix A.

B. Sentiment Engine

The sentiment engine that is used to classify the 9,476 articles extracted from Hegnar.no combines the sub-fields of Artificial Intelligence known as Computational Linguistics and

Machine Learning. First, each article is converted to a numerical vector such that each vector element corresponds to an attribute of the article (like length of title, number of times a ticker is mentioned in the text, whether the article has been published while the OSE is open etc.). Many of these attributes have been extracted using a 10,000-word sentiment lexicon tailored for the Norwegian financial news domain. Then a manually annotated (i.e. tagged) training set of 990 articles are used by the machine learning classification model J48 decision trees[7] to learn appropriate mapping from input article vector to output sentiment classification (positive, neutral, or negative). The sentiment engine achieved a precision of $\sim$70%, which is in line with the state-of-the-art in this subjective, imprecise task of sentiment classification (Balahur et al., 2010; Mourad and Darwish, 2013). For any interested reader, the methods used in this system are described in elaborate detail in Njølstad et al. (2014a) and Njølstad et al. (2014b).

C. Sentiment Index

In order to be able to study the interaction between the aggregate market-wide sentiment and firm-specific news sentiments, an index measuring this former "general market mood" was developed. Selecting among the potential variables reviewed by Baker and Wurgler (2007), guided by data availability, we chose the following: 1) number of IPOs in a period ($IPOC$), 2) average IPO return on first day of issue ($IPOR$), 3) retail fund flow ratio[8] ($RFFR$), 4) insider trade filing count[9] ($ITFC$), 5) total traded volume (in monetary value) on the OSE (VOL), 6) ratio of newly issued equity relative to bonds by stocks listed on the OSE (ETB), and 7) price-to-book value for the OSE combined (PB), for composition of the index. In the same fashion as the aforementioned authors, along with Brown and Cliff (2004), we use Principal Component Analysis[10] (Jolliffe, 1986) to aggregate the seven identified sentiment proxies into a single index, such that this single number explains the maximum of varation in these seven variables. This resulted in the equation: $senti_t = 0.907IPOC_t + 0.064IPOR_t + 0.350RFFR + 0.183ITFC_t + 0.101VOL_t + 0.084ETB_t + 0.019PB_t$. The index was computed using data with a monthly resolution. As an eyeball test to the composed aggregate market-wide sentiment index, we compare it to the historical development of the Oslo Stock Exchange All Shares Index (OSEAX). This is depicted in Figure 2.

[7]For a primer in machine learning, and hereunder decision trees, see Anderson et al. (1986) or Russell et al. (1995).

[8]Capital inflow divided by capital outflow for the period.

[9]All insider trades on the OSE are announced and registered through the named reporting portal NewsWeb.

[10]We use Weka's (www.cs.waikato.ac.nz/ml/weka) implementation of this statistical procedure.

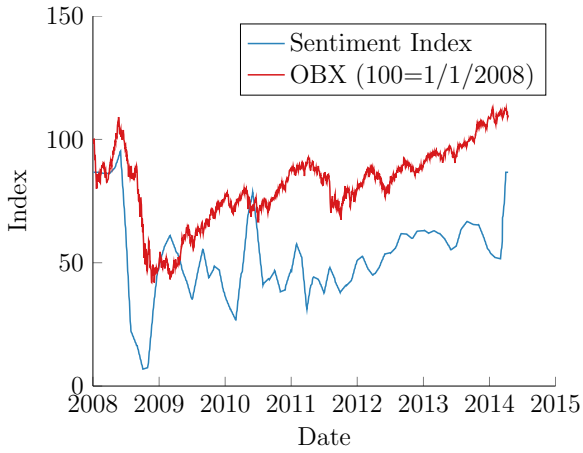


Figure 2. Historical development of Sentiment index and OSEAX. Both have been rebased to 100 at the 4^{th} of January 2008 (first day of trading this year).

From the figure it seems like the developed index correlates fairly well with OSEAX - it plunges upon the outbreak of the financial crisis late 2008 and resurges slowly, and in accordance with the OSEAX, after that.

III. Methodology

In this section we detail, and discuss, the statistic methodologies used in studying the relationships between the independent sentiment variables and dependent variables of stock price behavior. We first build ARCH regression models to regress stock price return, volume and order size with a number of sentiment and control variables. The ARCH regression model and variables used in these regressions are accounted for in Sections III.A and III.B, respectively. Secondly, we use hypothesis testing to compare coefficient outputs of regression models for different stocks to make statistical inference on the relationships between investor sentiments and stock price behavior. This is covered in Section III.C.

A. Regression models

The Ordinary Least Squares (OLS) linear regression model is popular in econometric analyses due to its desirable property of being the best linear unbiased estimator (BLUE) of coefficients in a linear regression model. However, this property, as with all desirable prop-

erties, only holds under certain assumptions, as stated by the Gauss-Markov Theorem[11] (Wooldridge, 2012). These assumptions cannot be satisfied in our case. In our dataset, residual errors are both correlated and of unequal variance; it exhibits time-varying volatility. Some periods display far more significant fluctuations than others. If we consider the graphs of stock return, traded monetary value, and average order size by ticker, visualized in Appendix A, it is clear from the vicissitude of these graphs that the variance, for neither of these three variables, can be assumed uncorrelated or constant over time. Fortunately, in our data, as is usually the general case for financial data (Campbell and Andrew, 1997), large and small such residual errors seem to occur in clusters; large returns, volumes, and order sizes are predominantly followed by more large returns, volumes, and order sizes, and vice versa. According to Wooldridge (2012), it is still possible to find efficient estimators in the presence of the heteroskedasticity that we have, but this requires knowing the form of this heteroskedasticity. If we allow the process describing the variance to be nonlinear and at the same time assume the process describing the mean to be linear (i.e. that the dependent and independent variables are linearly related), the autoregressive conditional heteroskedasticity (ARCH) model, for which Engle (1982) was awarded the Nobel price in Economics, can be used. This means that, for a model of order q, q lagged error terms with coefficients, $\alpha_i \epsilon_{t-i}^2,\ i \in \{1, 2, ..., q\}$, are added to the regression equation. These α coefficients are estimated by the model through the maximization of a log-likelihood function. For high orders, this can lead to convergence problems (Bollerslev, 1986).

We construct separate ARCH regression models for each of the stocks considered, since it is unreasonable to assume the variance of two completely different stocks to be the same[12], such as would be a prerequisite for using OLS and fixed-effects models. Mixed linear models[13] could have been used to handle this case of unequal variances between stocks, but proved infeasible for our purposes due to convergence problems, likely caused by high dimensionality. Since we are estimating three dependent variables, return, volume and order size, using daily, weekly and monthly data this equates to nine regression model formulations, which are listed in detail in Appendix B. These have again been estimated separately for each of the ten stocks, which in total gives 90 estimated regression models.

[11] The theorem states that the OLS estimator is BLUE when linear regression leads to residual errors that have expectation zero, are uncorrelated and have equal variances (Wooldridge, 2012).

[12] The blue chip stocks exhibit considerably less volatility in both return, volume and order size than white chips, as evident in Appendix A.

[13] Models allowing both fixed and random effects, like unequal variances of different stocks (Robinson, 1991).

A.1. Reverse causality and the endogeneity problem

For daily data, as described in full detail in Appendix B, we use the number of positive, neutral, and negative articles published on day t to predict the intraday return, $R_{t,i}$, traded monetary value, $T_{t,i}$, and order size, $O_{t,i}$, of ticker i within the same time interval, t. This has been done in lack of high-frequency data needed to perform analysis at the preferable lower level of granularity. Ideally, investor reaction to news should be traced down to the second-level, if not even lower. Hence, daily models will potentially suffer from so-called reverse causality - the independent variable could also be influenced by the dependent variable. In our dataset, using daily data, we cannot know for sure whether an article has caused the stock price of the article's target to increase, or if the price jump of a stock has caused a journalist to write an article with the stock as its target, seeking to explain what has happened. Possible solutions to this endogeneity[14] problem is to include lagged variables, which we have done for weekly and monthly data, in addition to using instrument variables (IV)[15]. We have also run the daily regressions with lagged independent variables (use yesterdays news to predict the stock price behavior today), which lead to an overall significant relationship between news and stock price behavior, but much less so than if considering daily data. Although this significance is intrinsically interesting, we have not reported these results as they are indicative of response reversion, moreso than the immediate reaction to news. In this paper we wish to just study this immediate reaction, and save this reversion for later endeavors at this time. Importantly, however, this significance from running daily lagged variables advocates that the significance found when considering same-day independent and dependent variables plausibly could be attributed to the true relationship between news and stock price behavior, and not just simply to the supposed endogenous relationship.

B. Variables

In this section we define and describe the variables used in our regression models. Since we are examining both daily, weekly and monthly data, the notion of time will vary between the regression models, and are expressed generally in this section. Precise notions of time (i.e. which time periods are considered when regressing daily, weekly, and monthly data) are given in full in Appendix B, for each of the three time horizons.

[14] A regression model is said to be endogenous when there is a correlation between an independent variable and the error term, which as it can be shown, is the case in the presence of reverse causality. This is a violation of linear regression models, like OLS, making coefficient estimates biased and inefficient and greatly limiting the statistical inference that can be drawn from such models (Wooldridge, 2012).

[15] Procedure where only the "exogenous" part of the variation in the independent variable, which necessarily needs to be related to the exogenous instrument variable, is included in the regression model through this instrument (Wooldridge, 2012).

B.1. Dependent Variables

$R_{t,i}$ Return of ticker i in time period t in percentage terms. If $cp_{t,i}$ is the closing price of ticker i at time t then $R_{t,i} = \frac{cp_{t,i} - cp_{t-1,i}}{cp_{t-1,i}}$.

$T_{t,i}$ Traded monetary value of ticker i in time period t relative to historical average over all n periods. If $tr_{t,i}$ is the traded monetary value of ticker i in time period t, then $T_{t,i} = \frac{tr_{t,i}}{\sum_{\tilde{t}=1}^{n} tr_{\tilde{t},i}}$. This relative measure is chosen to be able to compare the estimated β coefficients of this variable across stocks. The variable is quoted in percentage terms.

$O_{t,i}$ Average order size of ticker i in time period t relative to historical average over all n periods. If $os_{t,i}$ is the average order size of ticker i in time period t, then $O_{t,i} = \frac{os_{t,i}}{\sum_{\tilde{t}=1}^{n} os_{\tilde{t},i}}$. This relative measure is chosen to be able to compare the estimated β coefficients of this variable across stocks. The variable is also quoted in percentage terms.

B.2. Independent Variables

$pos_{t,i}$ Number of positive articles published with ticker i in its tag set in time period t.

$neu_{t,i}$ Number of neutral articles published with ticker i in its tag set in time period t.

$neg_{t,i}$ Number of negative articles published with ticker i in its tag set in time period t.

$sent_t$ Dummy variable equal to 1 if sentiment index value is greater than historical average ('Bull' sentiment) and 0 if below ('Bear' sentiment) in time period t. If $senti_t$ is the sentiment index in time period t then

$$sent_t = \begin{cases} 1 & \text{if } senti_t > \frac{1}{n}\sum_{\tilde{t}1}^{n} senti_{\tilde{t}} \\ 0 & \text{if } senti_t \leq \frac{1}{n}\sum_{\tilde{t}1}^{n} senti_{\tilde{t}} \end{cases}$$

$senti_t$ Sentiment index in time period t, computed according to the description in Section II.C.

h_t Dummy variable equal to 1 if time period t is a holiday (Easter, summer, or Christmas) 0 if it is not. This variable is only included when regressing daily data.

B.3. Control Variables

ro_t Return of OSEAX in time period t in percentage terms. If $cpo_{t,i}$ is the closing price of OSEAX at time t then, $ro_t = \frac{cpo_t - cpo_{t-1}}{cpo_t}$.

$\sigma_{t,t-q}$ Standard deviation of OSEAX return last q periods.

$nibor_t$ Norwegian InterBank Offered Rate (NIBOR) at time period t.

$oilprice_t$ The Brent Crude oil price in time period t.

$eurnok_t$ The EUR-NOK exchange rate in time period t.

$fsr_{t,i}$ Number of financial reporting announcements submitted via Newsweb (www.newsweb.no) by ticker i in time period t.

B.4. ARCH Variables

$\epsilon^2_{t-q,i}$ Squared standard error of estimate for the q^{th} previous time period and ticker i.

C. Hypothesis tests

As previously established, we construct separate ARCH regression models for each of the stocks considered. An aggregate stock model formulation, which for instance could include a dummy variable for stocks belonging to the blue chip category, would greatly simplify statistical inference. This would, as discussed in Section III.A, however violate the assumptions necessary for statistical inference due to unequal variances of stocks. In lack of such a model formulation, we need statistically sound ways of comparing coefficients from different models and comparing coefficients from different groups of models. For these purposes four hypothesis tests have been formulated, and these are detailed in Appendix E. These have all been implemented in Matlab and code listings are also included in Appendix F. We will now explain these four hypothesis tests in turn, at a high level.

C.1. Z-test of variable significance with Bonferroni correction

This test determines whether there is a significant relationship between an independent variable and the dependent variable across each stock model. Since this requires multiple comparisons, Bonferroni correction must be performed. In effect, this means that higher requirements on the z values associated with the β coefficients are needed to conclude that there is a significant relationship between the variables in place. The details of this test are given in Appendix E.E.1.

C.2. Wald test of variable difference

This test uses the premise that $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_k$ (Slotani, 1964) as a test statistic of β coefficient homogeneity between different models i.e. it determines whether the β coefficients output from different regression models are significantly different from each other. The elaborate details of this test, along with a simple proof of the arbitrarity in choice of contrast matrix, is depicted in Appendix E.E.2.

C.3. Two-sided Wald test of average variable difference

This test compares the average β coefficients of two groups of models to determine if they are significantly different from each other. This is done with a simple χ^2 test statistic (the square of a Z test statistic). The specifics of this test are described in Appendix E.E.3.

C.4. One-sided Wald test of average variable difference

This test also compares the average β coefficients of two groups of models to determine if they are significantly different from each other, but is one-sided as there is *a priori* reason to believe one to be smaller than the other. This is done with a simple Z test statistic. Full account of this test is given in Appendix E.E.4.

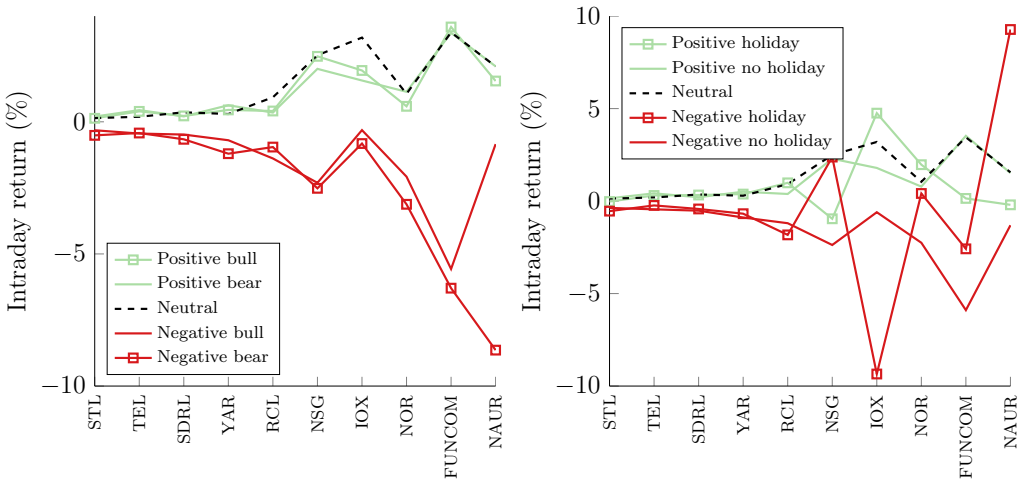
The results of running these hypothesis tests are summarized in the tables in Appendix D. These will be discussed in detail in the next section.

IV. Results and Discussion

The results from running the nine regression models for the ten different stocks, as detailed in Appendix A, are depicted in Appendix C. These show the estimated β coefficients of all included variables, for each of the ten stocks being studied, in the models trying to predict daily return (Table 19), daily traded monetary value (Table 20), daily order size (Table 21), average weekly return (Table 22), weekly traded monetary value (Table 23), average weekly order size (Table 24), monthly return (Table 25), monthly traded monetary value (Table 26), and, lastly, average monthly order size (Table 27). We will now present and discuss these outputs, along with the performed hypothesis tests (summarized in Appendix D) to make inference on the promises of this paper: 1) whether the relation between news sentiment and stock price behavior differs between blue chip and white chip stocks, 2) whether there exist firm-idiosyncratic news sentiment reactions, 3) whether there exist aggregate market-wide sentiment and firm-specific news interactions, and 4) whether there exist aggregate market-wide sentiment and seasonality interactions impacting news sentiment reactions. The discussion will follow the same structure as the regression output: daily data will be assessed first (in Section IV.A), then weekly data will be treated (Section IV.B), and, lastly monthly data (Section IV.C) will be evaluated.

A. Daily data

The regression models using daily data predict significant relationships between positive, neutral, and negative news sentiments and all of the three dependent variables intraday return, traded monetary value, and order size. This is evident in the first column of Table VIII, IX, and X, who showing results of running the aforementioned hypothesis tests. Furthermore, there exist firm-idiosyncratic reactions to positive, neutral and negative sentiment news (except for negative articles and order size), as evident in the second column of the same three tables. Moreover, when considering the differences between blue chip and white chip stocks (the results of running the hypothesis test that tests this are presented in the third column in the same tables) it is clear that these two groups, when aggregated, react significantly different. This is further clear inspecting the graphs depicted in Figure 3a and 3b. These graphs show the estimated β coefficients for predicting daily return, with sentiment index included in the former and seasonality interaction in the latter, with the publication of a positive, neutral, and negative article for the ten different stocks. The x-axis has been sorted by average traded volume (liquidity) such that the leftmost five stocks are the most liquid (blue chips) and the rightmost five stocks are the most illiquid (white chips). Both blue chips and white chips generally exhibit positive returns following positive news, and the converse for negative news. However, white chips seem to react more strongly to both positive and negative news. This is consistent with what one would expect - more information asymmetry is associated with illiquid stocks and any new piece of information (or simply sentiment if news contain no new information, as most often is the case) is, hence, more likely to cause changes in an investors sentiment towards the stock and induce a reaction. Our results are consistent with the Sentiment Seesaw model presented by Baker and Wurgler (2007); the authors predict the return of illiquid stocks to react more positively to positive sentiment and more negatively to negative news than liquid stocks.

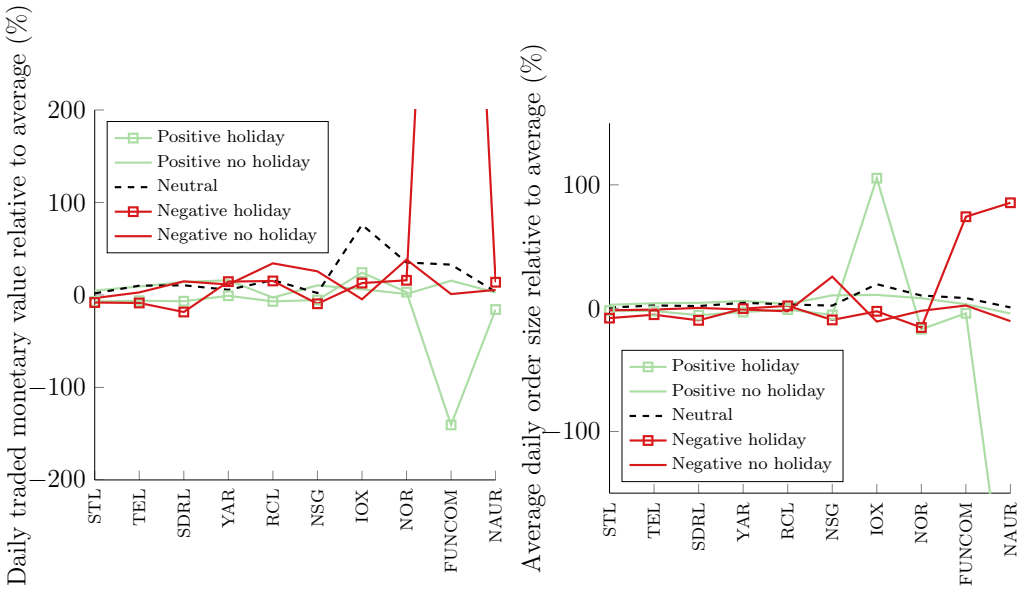


(a) Estimated β coefficients for predicting daily return in interaction with sentiment index

(b) Estimated β coefficients for predicting daily return holiday in interaction with holiday dummy

Figure 3. β coefficient graphs predicting return using daily data

Turning to interactions between the aggregate market-wide sentiment index and news flow, it is clear from the same tables in Appendix C (Table VIII, IX, and X) that it in general has a significant impact on stock price behavior (return upon positive news, volume, and order size for both types of signed news). Whether this interaction is different for blue chips and white chips is, however, less clear from the hypothesis tests. Lastly, assessing the interactions between seasonality (holiday) and news sentiment, it is further apparent from the hypothesis tests, in the same tables, that this interaction has a significant impact on both return, volume, and order size. Again, the distinction between the two groups of stocks cannot be established. Considering Figure 4a and 4b, however, it seems like the disperion, when seasonality is included, is much greater for white chips than blue chips. The direction of this effect, however, remains unclear. This seems to be the reason why the hypothesis tests in column three of Table VIII, IX, and X fail to reject H_0, as they only consider average difference between the two groups.



(a) Estimated β coefficients for predicting daily traded monetary value in interaction with holiday

(b) Estimated β coefficients for predicting daily order size in interaction with holiday

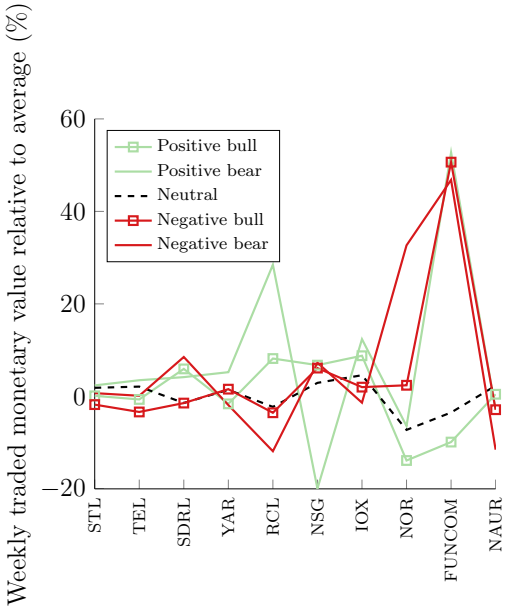
Figure 4. β coefficient graphs predicting traded monetary value and average order size using daily data

As discussed in Section III.A.1, the daily data has a resolution that potentially makes it liable to the endogeneity problem, caused by the presence of reverse causality (i.e. not possible to distinguish between stock price behaviors driven by news and, conversely, journalist publications driven by observed stock price developments). Hence, we conservatively only use the findings on the bases of these data to draw tentative conclusions; it justifies, and maps out the directions of further, more granular investigations. In-depth analyses using high-frequency data, not subject to the reverse causality caused endogenity problem, is hence a natural next step.

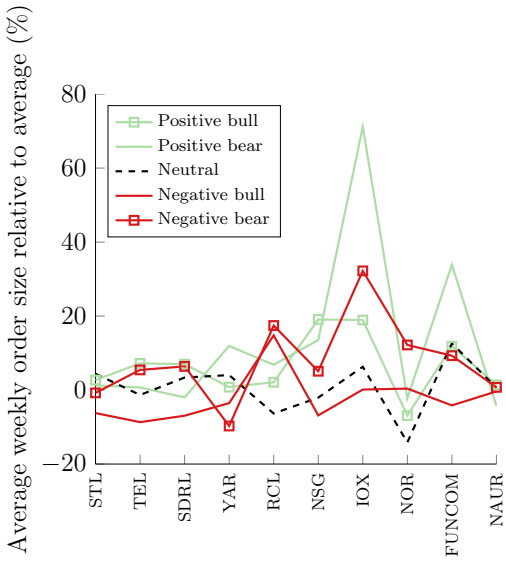
B. Weekly data

The regression models using weekly data predict significant relationships between positive, neutral, and negative news sentiments and traded monetary value. Only positive and neutral news (and not news classified as negative) significantly influence average order size. This

is evident in the hypothesis test results reported in the first column of Table XI and XII[16]. Considering the next column, assessing whether there exist firm-idiosyncratic reactions to positive, neutral, and negative sentiment news, it shows that positive and negative news impact on traded monetary value differ significantly between firms. Only positive and neutral news impact on order size differ significantly. There is not enough evidence to infer, however, that these found impacts vary significantly between white chips and blue chips, which again, probably is caused by the hypothesis tests working with group averages. Figure 5a and 5b show the estimated β coefficients for predicting weekly traded monetary value and average order size, respectively, with the publication of a positive, neutral, and negative article in interaction with the aggregate market-wide sentiment index for the ten different stocks. Visual inspection of these reveals that impact of news fluctuates much more among white chips (five rightmost stocks) than blue chips (five leftmost), although the directionality is still somewhat ambiguous.



(a) Estimated β coefficients for predicting weekly traded monetary value



(b) Estimated β coefficients for average predicting weekly order size

Figure 5. β coefficient graphs predicting traded monetary value and average order size using weekly data in interaction with the sentiment index

[16]Note that a table for hypothesis tests of regressing return using weekly data has been omitted from the appendix since only one significant relationship was found, that between publication of negative news and return as evident in Table 22, resulting in a near-empty table.

Turning to interactions between the aggregate market-wide sentiment index and news flow, it is clear from the tables in Appendix C (Table XI and XII) that both positive and negative news, in interaction with the aggregate market-wide sentiment, influences traded monetary value and that this effect is significantly different across the different firms. The same is only true for positive news when predicting order size. It is further clear from the hypothesis tests (clear from column three in Table XI and XII) that none of these interactions are significantly different for blue chips and white chips.

C. Monthly data

When examining monthly data, the regression models predict significant relationships between positive, neutral, and negative news sentiments and traded monetary value only (i.e. no significance is found from regressing return nor order size). This is evident in the hypothesis test results reported in the first column of Table XIII[17]. Considering the next column in the same table, assessing whether there exist firm-idiosyncratic reactions to positive, neutral, and negative sentiment news, it is clear that that the impact of all news types (positive, neutral, as well as negative) on traded monetary value differ significantly between firms. However, there is not enough evidence to infer that this impact varies significantly between white chips and blue chips. From Figure 6, however, it seems clear that white chip stocks lead to more violent reactions to news flow. As remarked in the two previous sections, one can speculate whether the lack of statistically significant differences is due to the simple averages of the two groups being compared in the hypothesis test.

[17]Note that tables for hypothesis tests of regressing return and order size using monthly data has been left out since no significance was found for these two independent variables, resulting in empty tables.

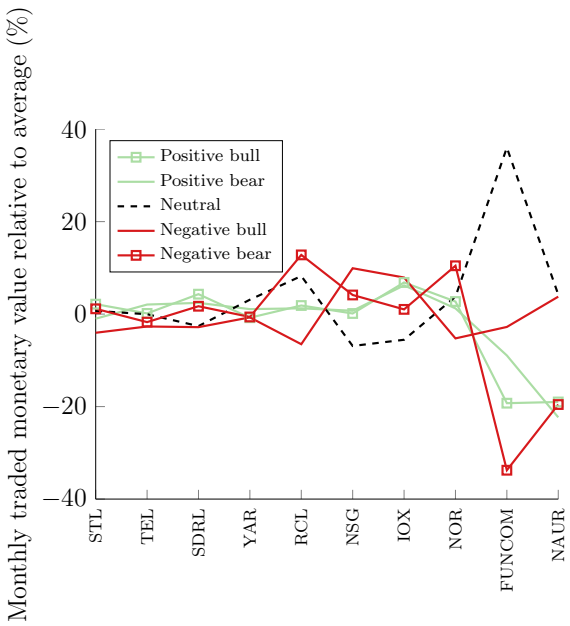


Figure 6. Estimated β coefficients for predicting monthly traded monetary value in interaction with the sentiment index

Considering the interactions between the aggregate market-wide sentiment index and news flow, it is clear from the Table XIII that both positive, neutral, and negative news, in interaction with the aggregate market-wide sentiment, influences traded monetary value and that this effect is significantly different across the different firms. It is also clear from the hypothesis tests (consulting column three in the same table) that none of these interactions are significantly different for blue chips and white chips.

In coda, the overall trend seems to be that published positive articles, ceteris paribus, increase volume (traded monetary value) while negative articles lead to a reduction of this same variable. The same seems to be the general proclivity for order size. Although there is evidence that publicity around certain single stocks lead to reduced average order size, the predominant tendency appears to be that this figure increases with news flow. Since institutional and other more "rational" investors will tend to drive order size up when trading, this observation, interestingly, seems to indicate that rational and irrational investors alike respond to the same news flow in a similar fashion. Furthermore, significant relationships between aggregate market-wide sentiment index and news impact on both volume and order have been unveiled. The lack of systematic differences between the average effect of blue and white chips indicates that this is idiosyncratically determined by each examined stock. The

graphs discussed in this section, however, seems to indicate that white chips have stronger idiosyncratic reaction to this interaction. A deeper analysis with more stocks could, perhaps, lead to more refined aggregate conclusions on this relationship.

A number of other significant relationships have been predicted by the regression models estimated using daily data, like that of news flow and stock price return with and without seasonal interactions. However, we are not confident enough in these findings, given the endogeneity problem supposedly caused by reverse causality, to draw more than tentative conclusions from these. This does indicate, nonetheless, without being able to determine the direction of causality, that there is a very strong relationship between news flow and return. Coupled with several individual stocks exhibiting stand-alone significant (alas not Bonferroni corrected) predictive relationships between news flow and return, this justifies further investigation with high-frequency intraday data. The fact that negative articles significantly impacts weekly return, as delineated in Section IV.B, even though the same significance is found for positive and neutral news articles, adds support to the supposition that there is indeed a causal relationship between news flow and short term stock return .

V. Conclusion

In this paper we have used a developed sentiment engine classifying firm-specific news as positive, neutral, or negative and a sentiment index to investigate the interaction impact of these on stock price behavior of shares listed on the OSE. For this investigation we have used Autoregressive conditional heteroskedasticity (ARCH) models and formulated several hypothesis tests for making statistical inference. We have found that publication of articles classified as positive by our sentiment engine generally lead to increases in traded volume, while negative articles have the opposite effect. The same appears to be the tendency for order size. In regressing return, only negative articles were found significant and the publication of these generally lead to reduced subsequent returns. The interaction between news articles and the aggregate market-wide sentiment index, which is the main focus of the paper, is also found to be statistically significant. Although the sign of this effect seems to be firm-idiosyncratic, our analysis indicates that white chips' reactions are stronger than that of blue chips.

Due to data availability being limited to low resolution intraday data, and suspicion of the reverse causality caused endogeneity problem, we only draw tentative conclusion on the relationship between the named interaction and return. With this, we advocate further analyses with high-frequency data. This represents the most natural venue for future work. Additionally, broadening our investigation to include more stocks, if not all, of the stocks

listed on the OSE could lead to stronger, more general conclusions. And lastly, comparing this relationship between news sentiments and stock price behavior across stock exchanges could also be of great interest, as this could shed some light on which exchanges should be the target of news sentiment trading strategies.

Appendix A. Descriptive statistics

Appendix A.1. Tables of summary statistics

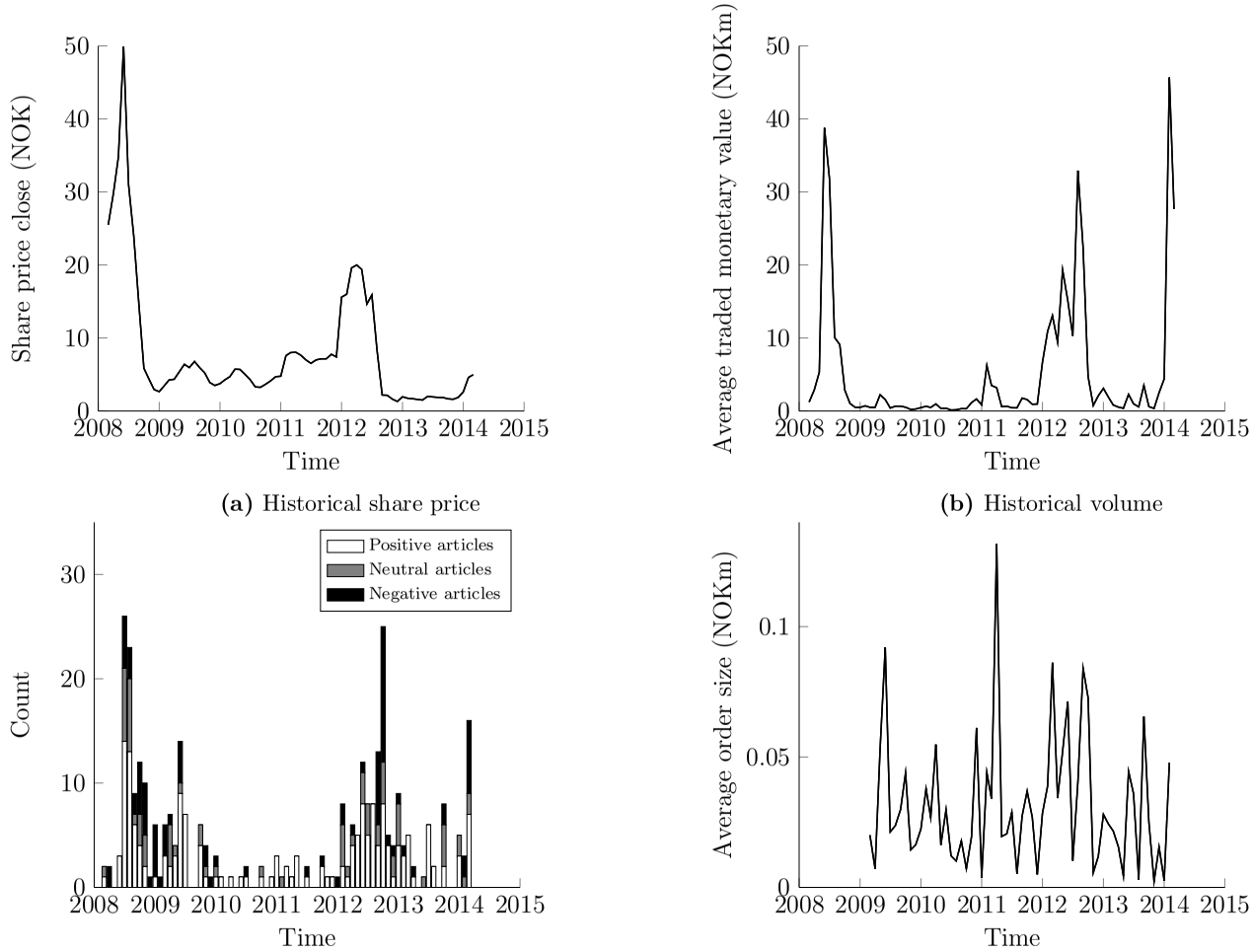
Ticker	Market value [NOKm] (% of OSE)	Average Daily Traded Monetary Value [NOKm]	Standard Deviation Daily Traded Monetary Value [NOKm]	Average Daily Order size [NOK]	Standard Deviation Daily Order size [NOK]
FUNCOM	346.5 (0.02)	5.6	13.6	15869	11667
IOX	317.4 (0.02)	1.7	5.2	20619	30186
NAUR	882.01 (0)	0.8	4.7	1207	3108
NOR	147.15 (0.06)	10.7	22.5	38350	53707
NSG	1026.1 (0.05)	24.9	36.8	24688	13710
RCL	65483.7 (3.53)	129	126	77508	39071
SDRL	92370.7 (5)	363	224	85424	33559
STL	539691.7 (29.13)	1130	752	173060	70408
TEL	195480 (10.55)	352	219	89049	25207
YAR	72753.57 (3.93)	627	485	91131	37301

Table III Market value, volume, and order size statistics by ticker

Ticker	Return from 01.01.2008 to 14.07.2014 [%]	Average intraday return[%]	Standard deviation intraday return [%]	Average weekly return[%]	Standard deviation weekly return	Average monthly return[%]	Standard deviation monthly return[%]
FUNCOM	-0.85	0.05	5.9	0.24	11.6	1.67	27.53
IOX	-0.96	-0.05	5.5	-0.31	11.2	-0.64	23.74
NAUR	-1	-0.17	6.4	-0.6	13.3	-1.7	31.23
NOR	-0.99	-0.19	4.8	-0.84	10.3	-3.05	20.43
NSG	-0.90	-0.03	4.8	-0.08	10.5	-0.34	20.77
RCL	0.47	0.08	3.4	0.52	8.2	1.79	16.14
SDRL	0.5	0.07	2.9	0.41	6.2	1.58	10.65
STL	0.03	0.02	2	0.12	3.9	0.35	6.28
TEL	0.03	0.03	2.2	0.13	4.2	0.8	9.94
YAR	-0.01	0.05	3	0.25	6.6	0.72	12.34

Table IV Return statistics by ticker

Appendix B.2. Firm-specific variables



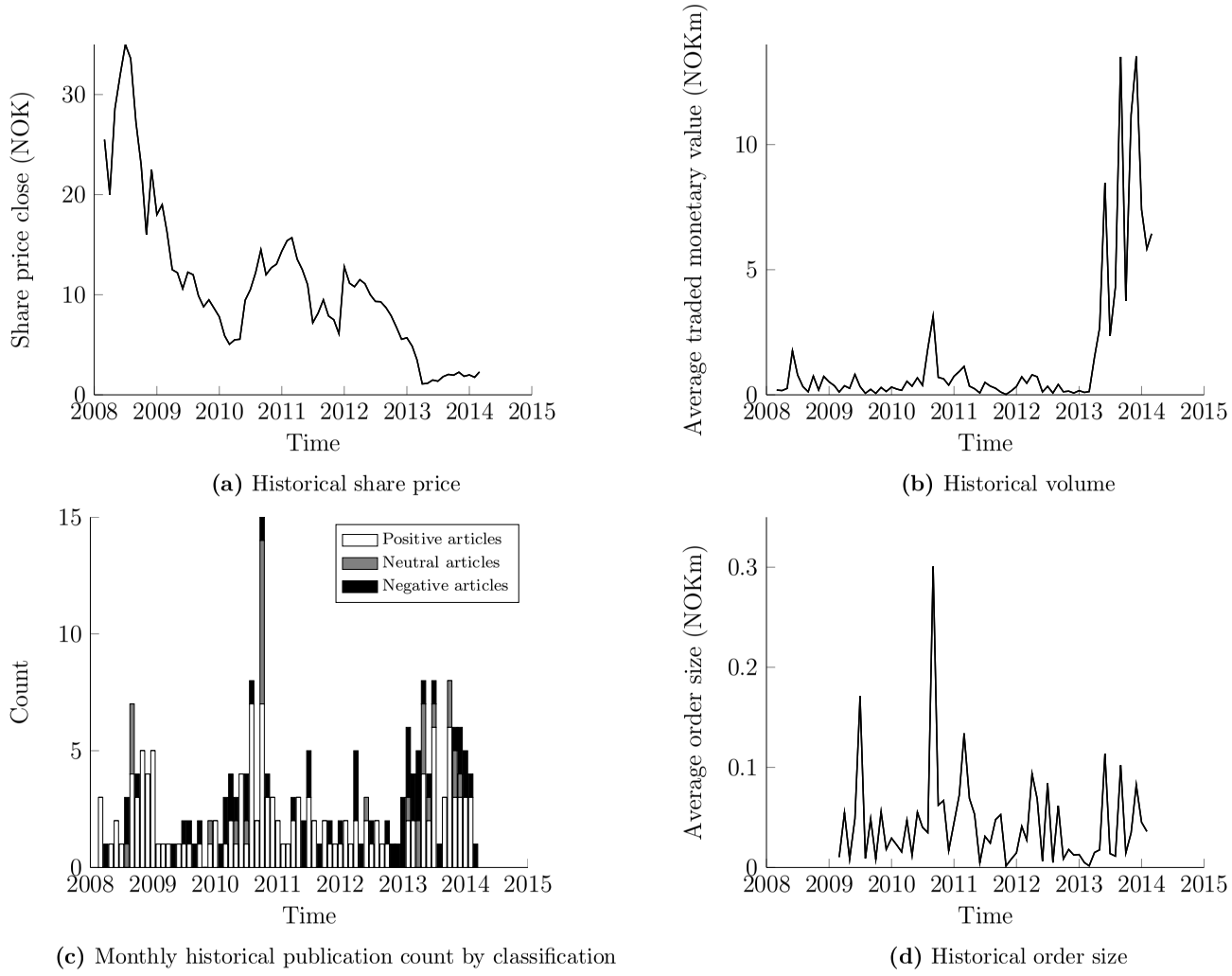
(a) Historical share price

(b) Historical volume

(c) Monthly historical publication count by classification

(d) Historical order size

Figure 7. Descriptive statistics of historical share price development and publication count for FUNCOM



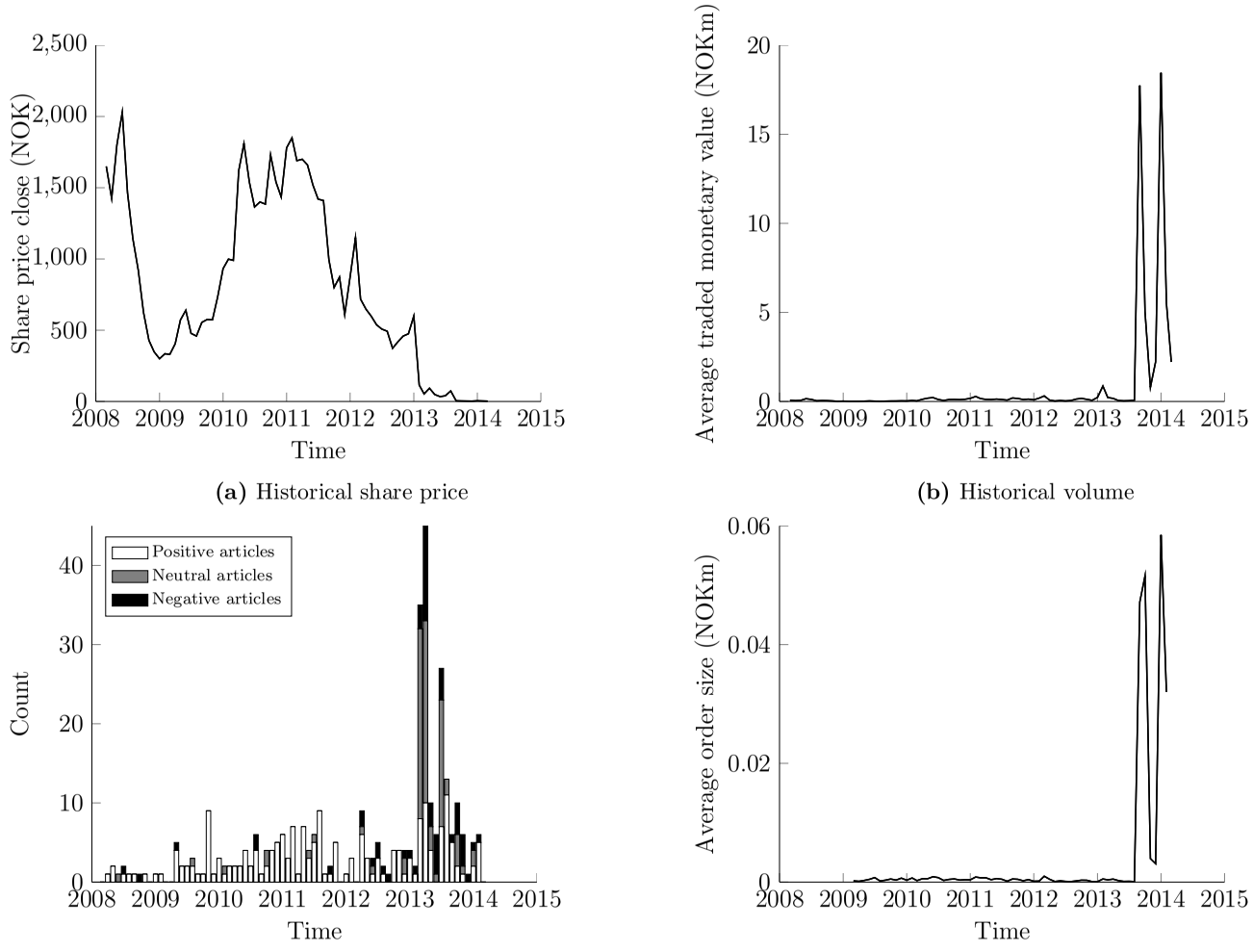
(a) Historical share price

(b) Historical volume

(c) Monthly historical publication count by classification

(d) Historical order size

Figure 8. Descriptive statistics of historical share price development and publication count for IOX



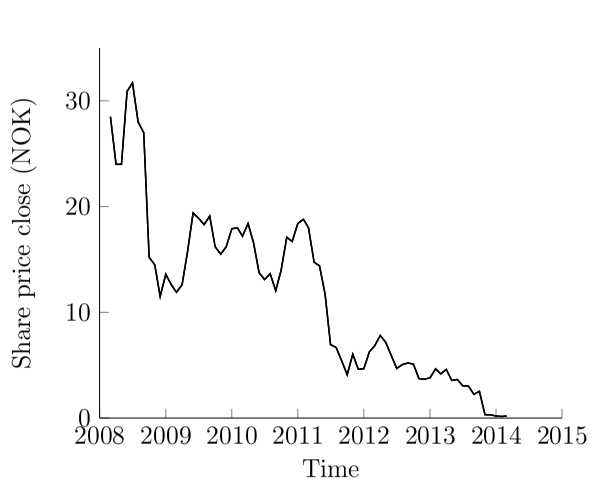
(a) Historical share price

(b) Historical volume

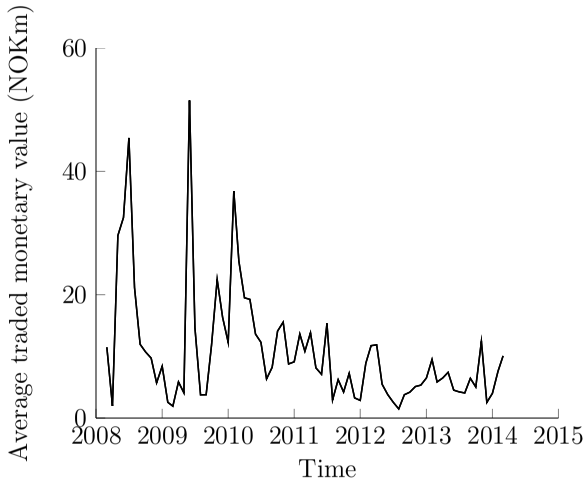
(c) Monthly historical publication count by classification

(d) Historical order size

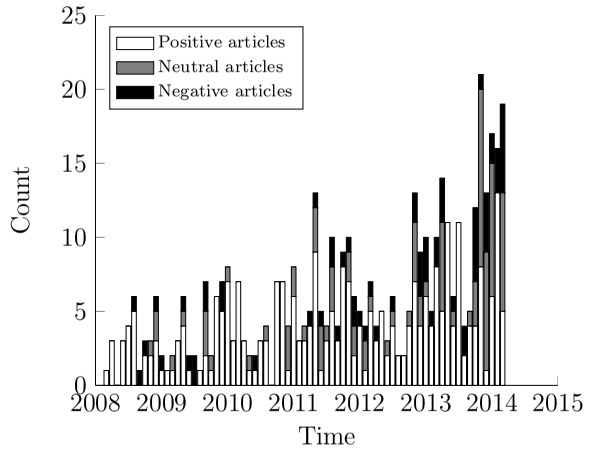
Figure 9. Descriptive statistics of historical share price development and publication count for NAUR



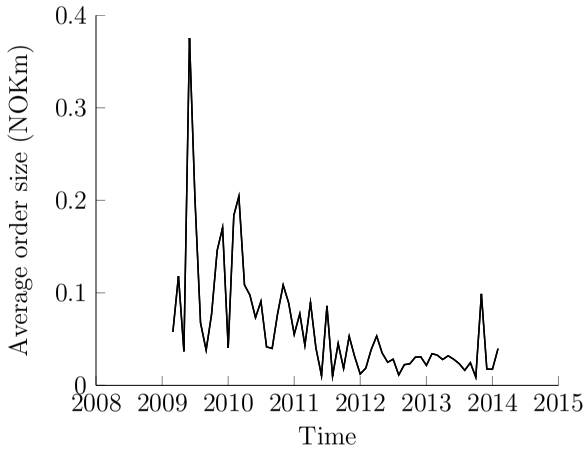
(a) Historical share price



(b) Historical volume

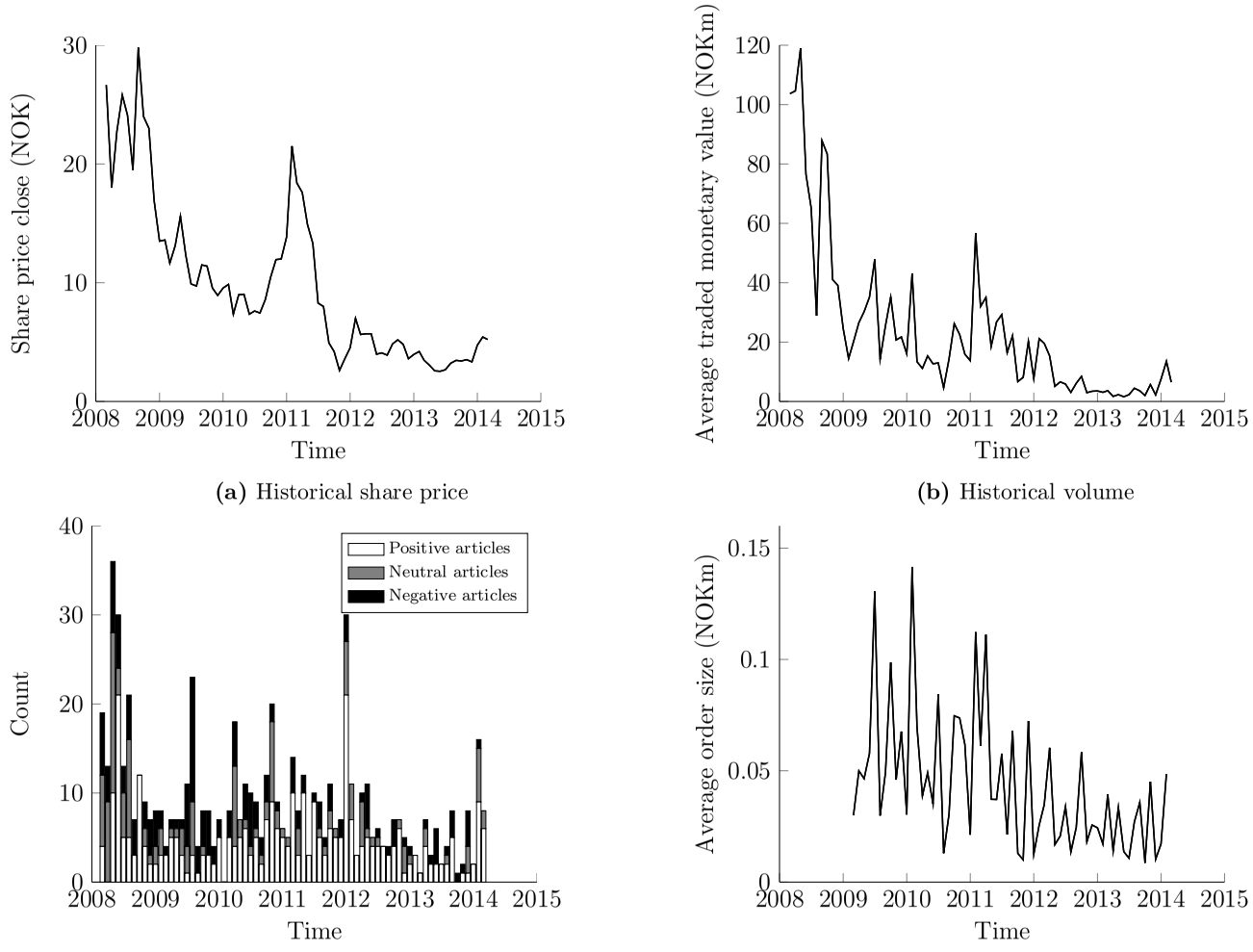


(c) Monthly historical publication count by classification



(d) Historical order size

Figure 10. Descriptive statistics of historical share price development and publication count for NOR



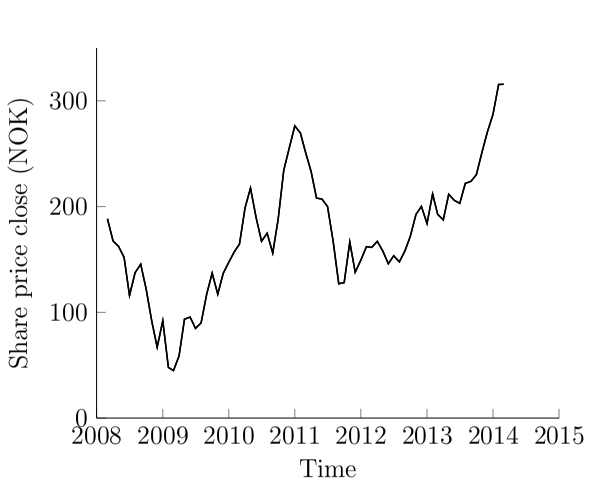
(a) Historical share price

(b) Historical volume

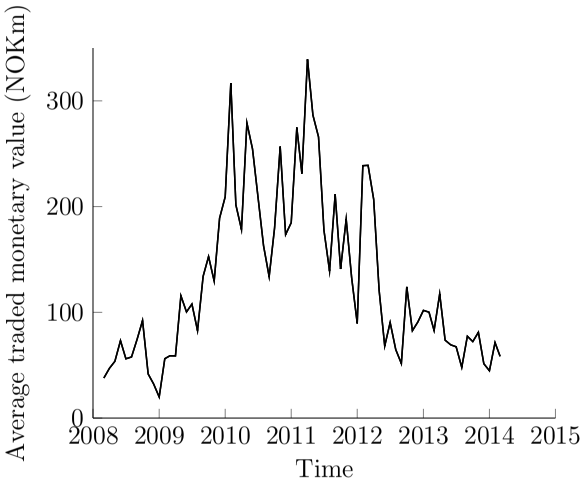
(c) Monthly historical publication count by classification

(d) Historical order size

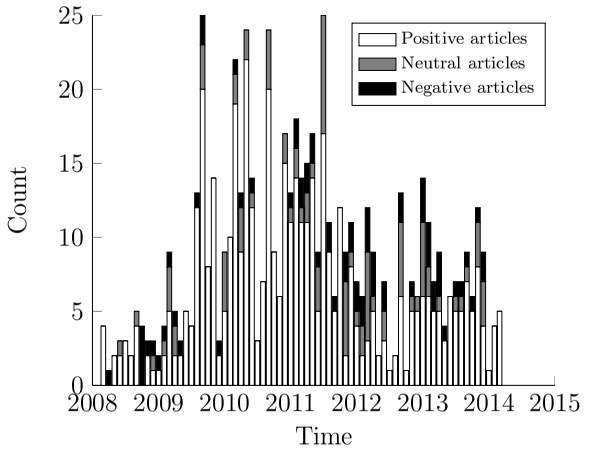
Figure 11. Descriptive statistics of historical share price development and publication count for NSG



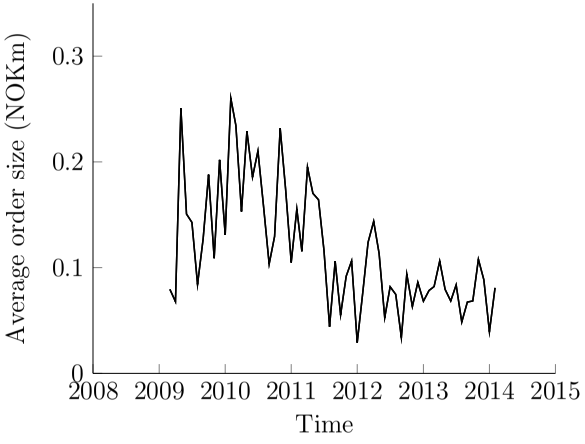
(a) Historical share price



(b) Historical volume

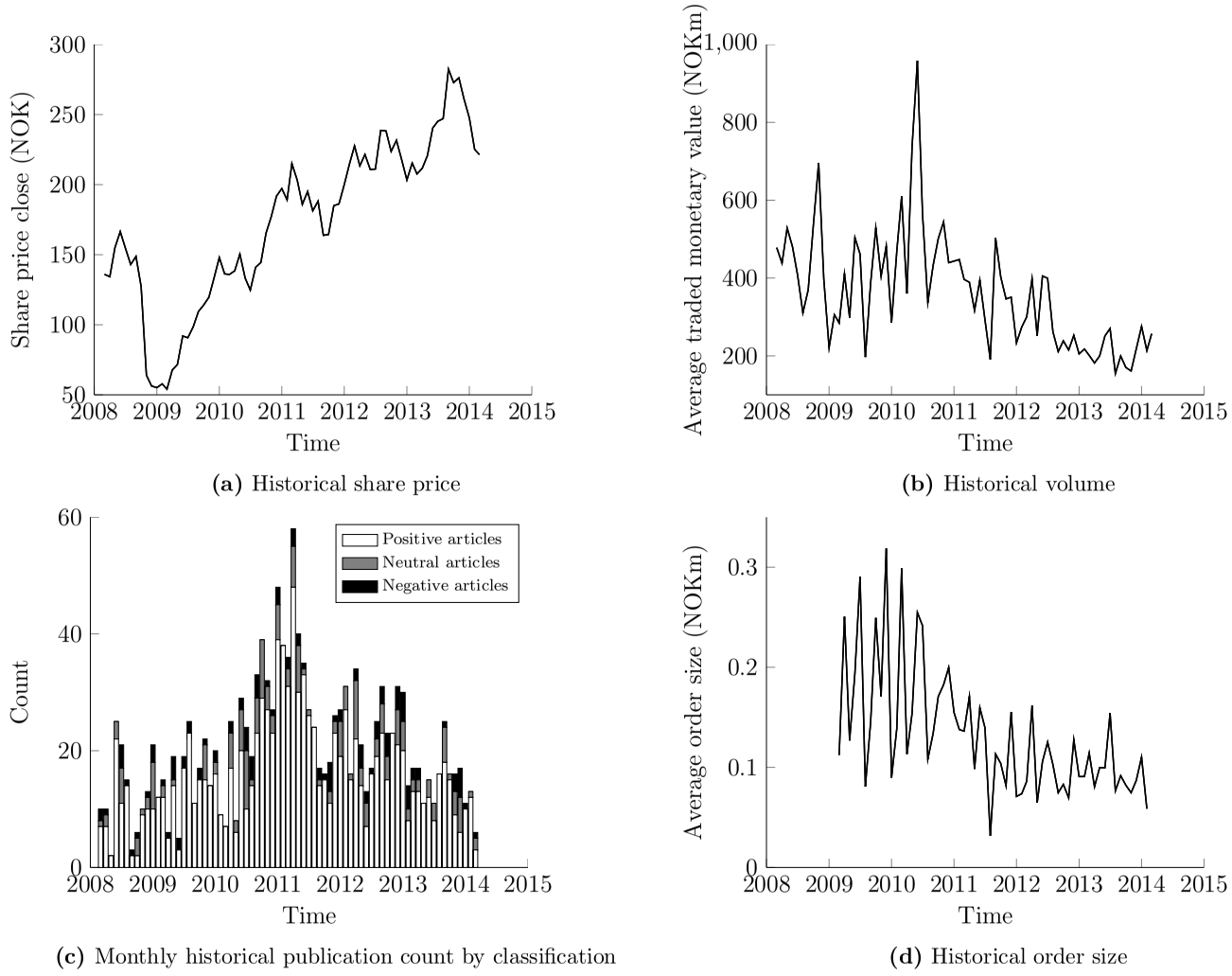


(c) Monthly historical publication count by classification



(d) Historical order size

Figure 12. Descriptive statistics of historical share price development and publication count for RCL



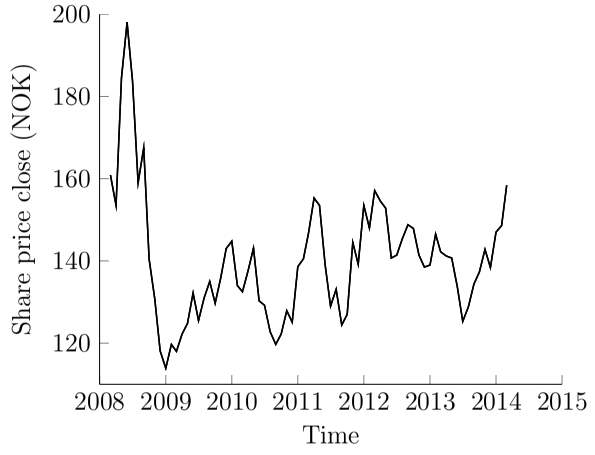
(a) Historical share price

(b) Historical volume

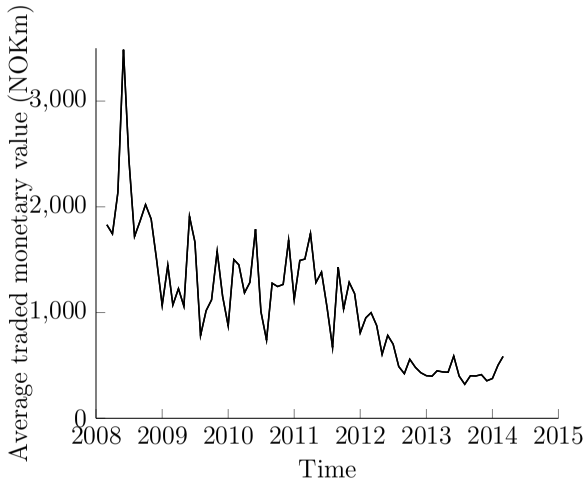
(c) Monthly historical publication count by classification

(d) Historical order size

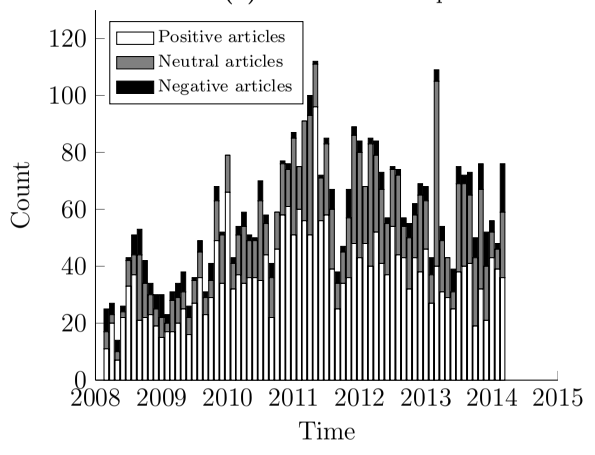
Figure 13. Descriptive statistics of historical share price development and publication count for SDRL



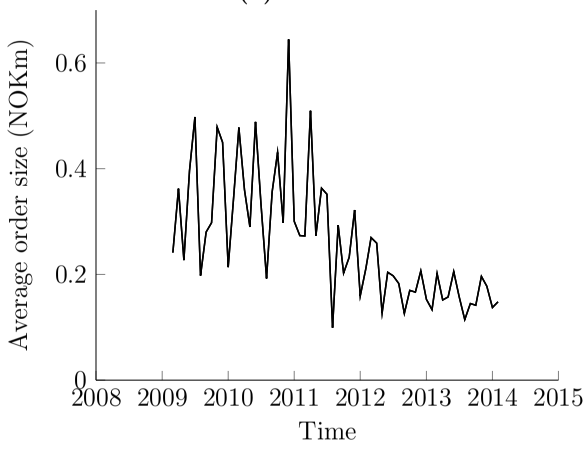
(a) Historical share price



(b) Historical volume

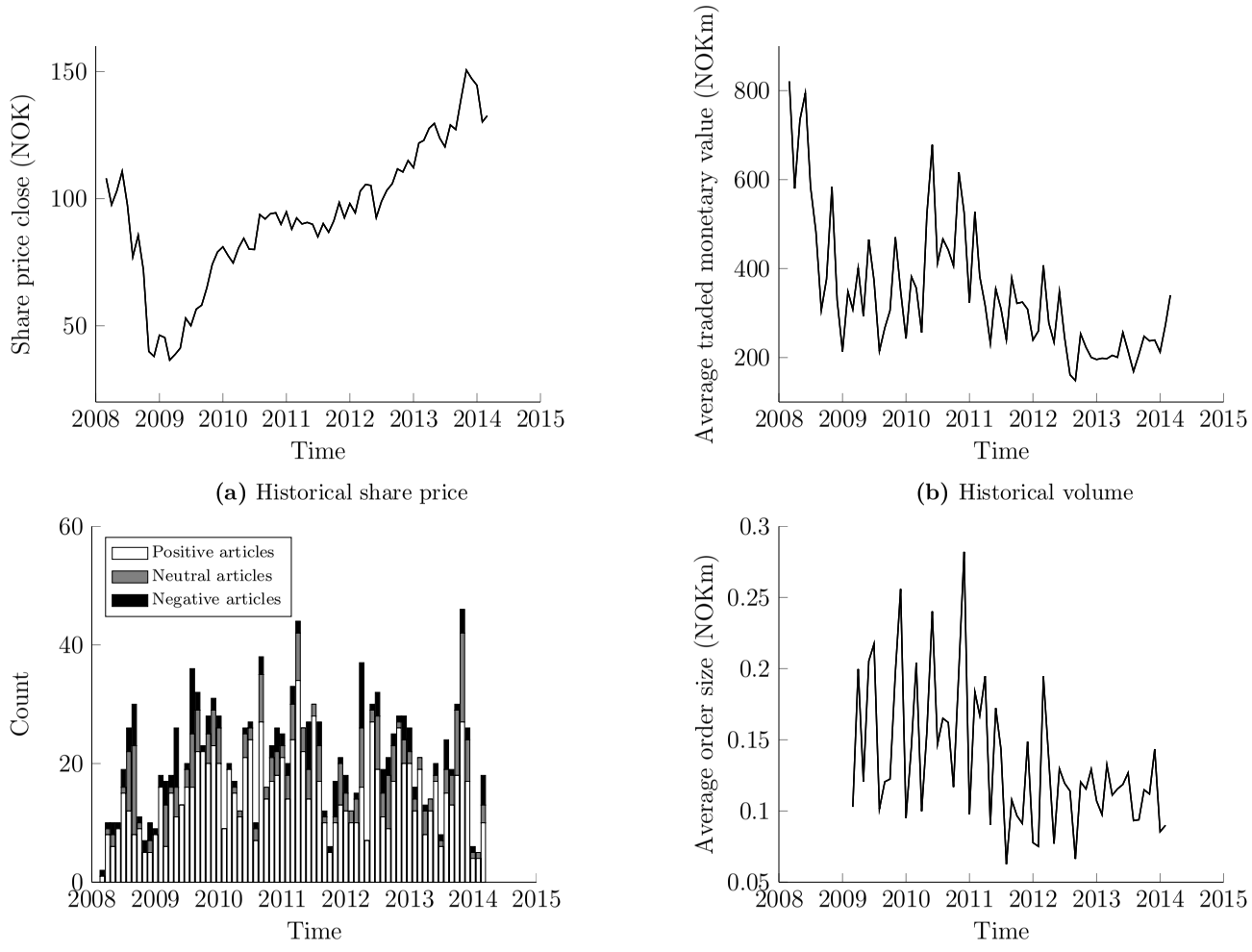


(c) Monthly historical publication count by classification



(d) Historical order size

Figure 14. Descriptive statistics of historical share price development and publication count for STL



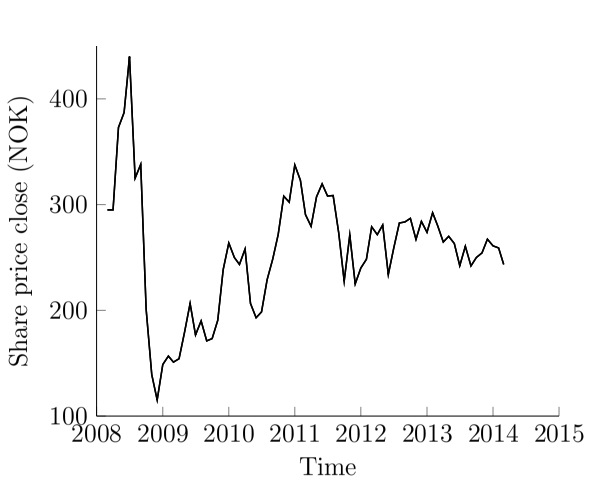
(a) Historical share price

(b) Historical volume

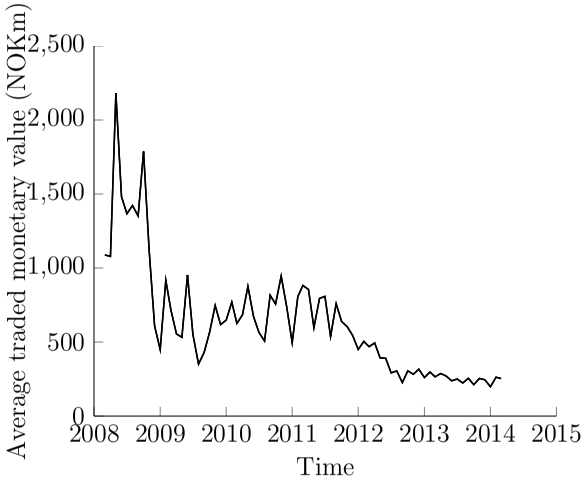
(c) Monthly historical publication count by classification

(d) Historical order size

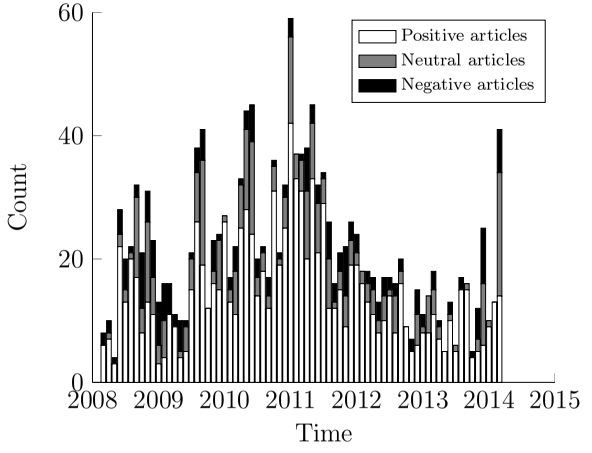
Figure 15. Descriptive statistics of historical share price development and publication count for TEL



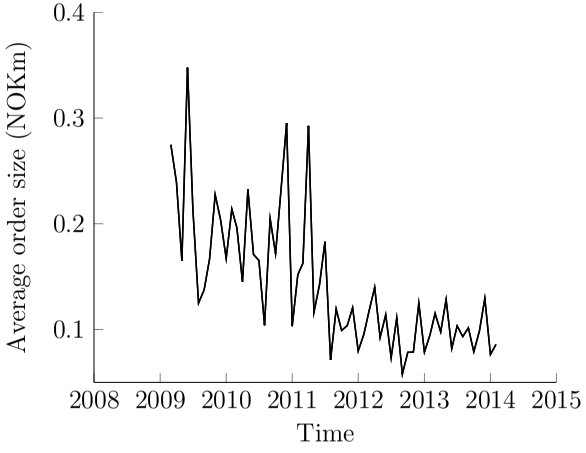
(a) Historical share price



(b) Historical volume

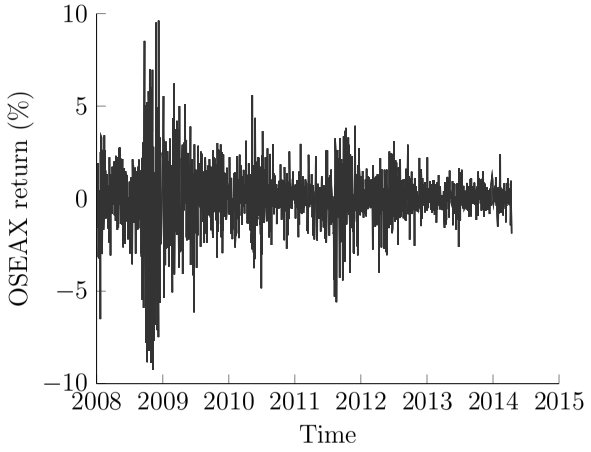


(c) Monthly historical publication count by classification

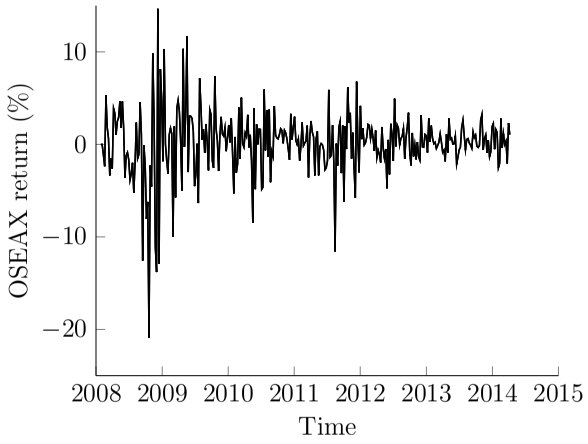


(d) Historical order size

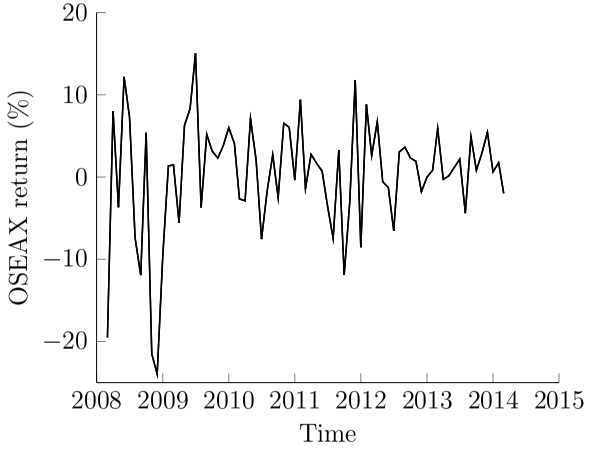
Figure 16. Descriptive statistics of historical share price development and publication count for YAR



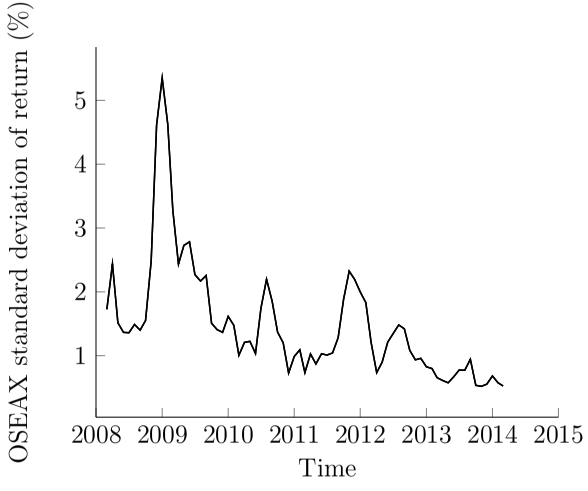
(a) Historical OSEAX intraday return



(b) Historical OSEAX weekly return

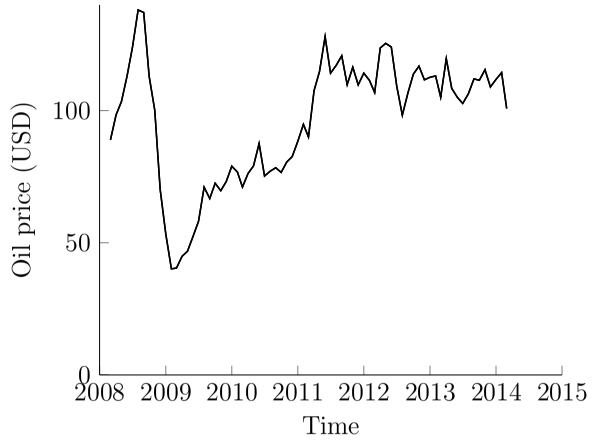


(c) Historical OSEAX monthly return

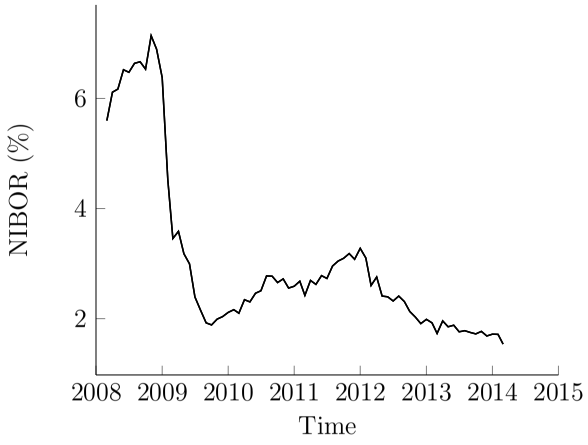


(d) Historical OSEAX standard deviation of return thirty days trailing

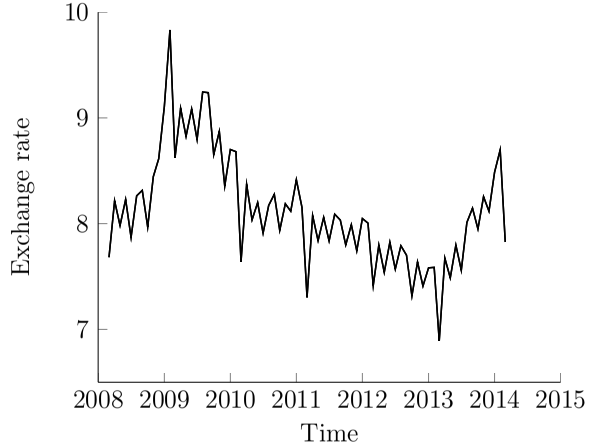
Figure 17. Descriptive statistics of historical developments of OSEAX control variables



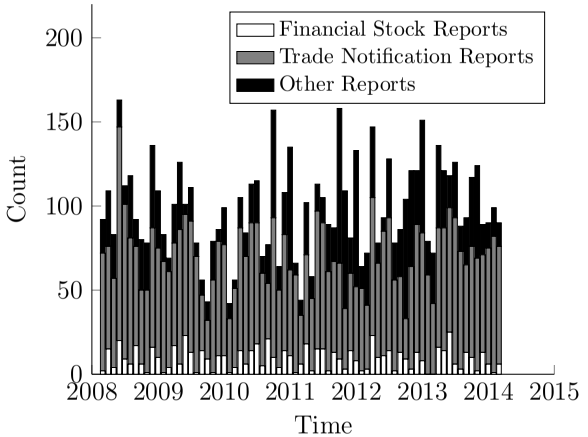
(a) Historical oil price



(b) Historical three-month NIBOR



(c) Historical EUR-NOK exchange rate



(d) Historical monthly stock announcement publication count

Figure 18. Descriptive statistics of historical developments of other control variables

Appendix B. Regression model formulations

In this section we account for the details of the regression models developed for this paper. All regressions were run with Autoregressive conditional heteroskedasticity (ARCH) models of varying order. This order was selected based on 1) whether significant increase in explanatory power from including an extra lagged term was achieved and 2) feasibility in convergence, which turned out to be an issue when many variables were included in the models.

Appendix B.1. Daily data

The variables used in the regression models on the daily data are listed and described in Table V. The specific regression equations used in predicting intraday return, traded monetary value, and order size are presented in the three following appendices.

Variable	Description
Dependent	
$R_{t,i}$	Intraday return of ticker i on day t [%].
$T_{t,i}$	Traded monetary value of ticker i on day t relative to historical daily average [%].
$O_{t,i}$	Order size of ticker i on day t relative to historical daily average [%].
Independent	
$pos_{t,i}$	Number of positive articles published with ticker i in tag set on day t.
$neu_{t,i}$	Number of neutral articles published with ticker i in tag set on day t.
$neg_{t,i}$	Number of negative articles published with ticker i in tag set on day t.
$sent_t$	Dummy variable equal to 1 if sentiment index value is greater than historical average (bull sentiment) and 0 if below (bear sentiment) at day t.
h_t	Dummy variable equal to 1 if holiday (Easter, summer, or Christmas) and 0 if no holiday at day t.
Control	
ro_{t-1}	Intraday return of OSEAX at trading day $t-1$.
$\sigma_{t,t-30}$	Standard deviation of OSEAX intraday return last thirty days.
$nibor_{t-1}$	Norwegian InterBank Offered Rate (NIBOR) at day $t-1$.
$oilprice_{t-1}$	The brent crude oil price at day $t-1$.
$eurnok_{t-1}$	The EUR-NOK exchange rate at day $t-1$.
$fsr_{t,i}$	Number of financial reporting announcements submitted via Newsweb (www.newsweb.no) by ticker i on day t.
ARCH	
$\epsilon^2_{t-q,i}, q \in \{1,2,...,5\}$	Squared standard error of estimate for day $t-q$ for ticker i.

Table V Variables used in regression of daily data with descriptions.

Appendix B.1.1. Intraday return

Intraday return for ticker i on day t, $R_{t,i}$, was regressed using an ARCH model of order 5 with independent variables expressing the number of published positive, neutral and negative articles on ticker i and day t and in interaction with the dummy variables sentiment index $sent_t$ and holiday h_t. A number of control variables were also added and those proving significant were kept. Formally:

$$R_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t,i} + \beta_{2,i} neu_{t,i} + \beta_{3,i} neg_{t,i} + \beta_{4,i} \sigma_{t,t-30} + \beta_{5,i} nibor_{t-1} + \beta_{6,i} fsr_{t,i} + \beta_{7,i}(pos_{t,i} \times sent_t) + \beta_{8,i}(neg_{t,i} \times sent_t) + \beta_{9,i}(pos_{t,i} \times h_t) + \beta_{10,i}(neg_{t,i} \times h_t) + \alpha_{1,i}\epsilon^2_{t-1,i} + \alpha_{2,i}\epsilon^2_{t-2,i} + \alpha_{3,i}\epsilon^2_{t-3,i} + \alpha_{4,i}\epsilon^2_{t-4,i} + \alpha_{5,i}\epsilon^2_{t-5,i} + u_{t,i}$$

The output from running the regression equation on the daily data is given in Table 19.

Appendix B.1.2. Daily traded monetary value

Traded monetary value for ticker i on day t, $T_{t,i}$, was regressed using a second-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i and day t and in interaction with the dummy variables sentiment index $sent_t$ and holiday h_t. Control variables were also added and those proving significant were kept. Formally:

$$T_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t,i} + \beta_{2,i} neu_{t,i} + \beta_{3,i} neg_{t,i} + \beta_{4,i} \sigma_{t,t-30} + \beta_{5,i} nibor_{t-1} + \beta_{6,i} fsr_{t,i} + \beta_{7,i}(pos_{t,i} \times sent_t) + \beta_{8,i}(neg_{t,i} \times sent_t) + \beta_{9,i}(pos_{t,i} \times h_t) + \beta_{10,i}(neg_{t,i} \times h_t) + \beta_{11,i} ro_{t-1} + \beta_{12,i} oilprice_{t-1} + \beta_{13,i} eurnok_{t-1} + \alpha_{1,i}\epsilon^2_{t-1,i} + \alpha_{2,i}\epsilon^2_{t-2,i} + u_{t,i}$$

The output from running the regression equation on the daily data is depicted in Table 20.

Appendix B.1.3. Average daily order size

Order size for ticker i on day t, $O_{t,i}$, was regressed using a first-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i and day t and in interaction with the dummy variables sentiment index $sent_t$ and holiday h_t. Control variables were also added and those proving significant were kept. Formally:

$$O_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t,i} + \beta_{2,i} neu_{t,i} + \beta_{3,i} neg_{t,i} + \beta_{4,i} \sigma_{t,t-30} + \beta_{5,i} nibor_{t-1} + \beta_{6,i} fsr_{t,i} + \beta_{7,i}(pos_{t,i} \times sent_t) + \beta_{8,i}(neg_{t,i} \times sent_t) + \beta_{9,i}(pos_{t,i} \times h_t) + \beta_{10,i}(neg_{t,i} \times h_t) + \beta_{11,i} ro_{t-1} + \beta_{12,i} oilprice_{t-1} + \beta_{13,i} eurnok_{t-1} + \alpha_{1,i}\epsilon^2_{t-1,i} + u_{t,i}$$

The output from running the regression equation on the daily data is shown in Table 21.

Appendix B.2. Weekly data

The variables used in the regression models on the weekly data are listed and described in Table VI. The specific regression equations used in predicting weekly return, traded monetary value, and order size are presented in the three following appendices.

Variable	Description
Dependent	
$R_{t,i}$	Weekly return of ticker i in week t [%].
$T_{t,i}$	Traded monetary value of ticker i in week t relative to historical weekly average [%].
$O_{t,i}$	Order size of ticker i in week t relative to historical weekly average [%].
Independent	
$pos_{t-1,i}$	Number of positive articles published with ticker i in tag set in week $t-1$.
$neu_{t-1,i}$	Number of neutral articles published with ticker i in tag set in week $t-1$.
$neg_{t-1,i}$	Number of negative articles published with ticker i in tag set in week $t-1$.
$sent_{t-1}$	Dummy variable equal to 1 if sentiment index value is greater than historical average (bull sentiment) and 0 if below (bear sentiment) in week $t-1$.
Control	
$\sigma_{t,t-4}$	Standard deviation of OSEAX intraday return last four weeks.
$nibor_{t-1}$	Norwegian InterBank Offered Rate (NIBOR) in week $t-1$.
$oilprice_{t-1}$	The Brent Crude oil price in week $t-1$.
$eurnok_{t-1}$	The EUR-NOK exchange rate in week $t-1$.
$fsr_{t,i}$	Number of financial reporting announcements submitted via Newsweb (www.newsweb.no) by ticker i in week t.
ARCH	
$\epsilon^2_{t-q,i}, q \in \{1,2\}$	Squared standard error of estimate for week $t-q$ for ticker i.

Table VI Variables used in regression of weekly data with descriptions.

Appendix B.2.1. Weekly return

Return in week t for ticker i , $R_{t,i}$, was regressed using a second-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i in week $t-1$ and in interaction with the dummy variable sentiment index $sent_{t-1}$. A number of control variables were also added and those proving significant were kept. Formally:

$$R_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t-1,i} + \beta_{2,i} neu_{t-1,i} + \beta_{3,i} neg_{t-1,i} + \beta_{4,i}\sigma_{t,t-4} + \beta_{5,i} nibor_{t-1} + \alpha_{1,i}\epsilon^2_{t-1,i} + \alpha_{2,i}\epsilon^2_{t-2,i} + u_{t,i}$$

The output from running the regression equation on the daily data is given in Table 22.

Appendix B.2.2. Weekly traded monetary value

Traded monetary value in week t for ticker i, $T_{t,i}$, was regressed using a second-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i in week $t-1$ and in interaction with the dummy variable sentiment index $sent_{t-1}$. A number of control variables were also added and those proving significant were kept. Formally:

$$T_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t-1,i} + \beta_{2,i} neu_{t-1,i} + \beta_{3,i} neg_{t-1,i} + \beta_{4,i}\sigma_{t,t-4} + \beta_{5,i} nibor_{t-1} + \beta_{6,i}(pos_{t-1,i} \times sent_{t-1}) + \beta_{7,i}(neg_{t-1,i} \times sent_{t-1}) + \beta_{8,i} + oilprice_{t-1} + \beta_{9,i} eurnok_{t-1} + \beta_{10,i} fsr_{t,i} + \alpha_{1,i}\epsilon^2_{t-1,i} + \alpha_{2,i}\epsilon^2_{t-2,i} + u_{t,i}$$

The output from running the regression equation on the daily data is depicted in Table 23.

Appendix B.2.3. Average weekly order size

Order size in week t for ticker i, $O_{t,i}$, was regressed using a second-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i in week $t-1$ and in interaction with the dummy variable sentiment index $sent_{t-1}$. A number of control variables were also added and those proving significant were kept. Formally:

$$O_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t-1,i} + \beta_{2,i} neu_{t-1,i} + \beta_{3,i} neg_{t-1,i} + \beta_{4,i} \sigma_{t,t-4} + \beta_{5,i} nibor_{t-1} + \beta_{6,i}(pos_{t-1,i} \times sent_{t-1}) + \beta_{7,i}(neg_{t-1,i} \times sent_{t-1}) + \beta_{8,i} + oilprice_{t-1} + \beta_{9,i} eurnok_{t-1} + \beta_{10,i} fsr_{t,i} + \alpha_{1,i}\epsilon^2_{t-1,i} + \alpha_{2,i}\epsilon^2_{t-2,i} + u_{t,i}$$

The output from running the regression equation on the daily data is shown in Table 24.

Appendix B.3. Monthly data

The variables used in the regression models on the monthly data are listed and described in Table VII. The specific regression equations used in predicting monthly return, traded monetary value, and order size are presented in the three following appendices.

Variable	Description
Dependent	
$R_{t,i}$	Monthly return of ticker i in week t [%].
$T_{t,i}$	Traded monetary value of ticker i in month t relative to historical monthly average [%].
$O_{t,i}$	Order size of ticker i in month t relative to historical monthly average [%].
Independent	
$pos_{t-1,i}$	Number of positive articles published with ticker i in tag set in month $t-1$.
$neu_{t-1,i}$	Number of neutral articles published with ticker i in tag set in month $t-1$.
$neg_{t-1,i}$	Number of negative articles published with ticker i in tag set in month $t-1$.
$sent_{t-1}$	Dummy variable equal to 1 if sentiment index value is greater than historical average (bull sentiment) and 0 if below (bear sentiment) in month $t-1$.
Control	
ro_{t-1}	Return of OSEAX in month $t-1$.
σ_{t-1}	Standard deviation of OSEAX intraday return last month.
$nibor_{t-1}$	Norwegian InterBank Offered Rate (NIBOR) in month $t-1$.
$oilprice_{t-1}$	The Brent Crude oil price in month $t-1$.
$eurnok_{t-1}$	The EUR-NOK exchange rate in month $t-1$.
$fsr_{t,i}$	Number of financial reporting announcements submitted via Newsweb (www.newsweb.no) by ticker i in month t.
ARCH	
$\epsilon^2_{t-1,i}$	Squared standard error of estimate for month previous month and ticker i.

Table VII Variables used in regression of monthly data with descriptions.

Appendix B.3.1. Monthly return

Return in month t for ticker i, $R_{t,i}$, was regressed using a fist-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i in month $t-1$ and in interaction with the dummy variable sentiment index $sent_{t-1}$. A number of control variables were also added and those proving significant

were kept. Formally:

$$R_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t-1,i} + \beta_{2,i} neu_{t-1,i} + \beta_{3,i} neg_{t-1,i} + \beta_{4,i}\sigma_{t-1} + \beta_{5,i} ro_{t-1} + \beta_{6,i} nibor_{t-1} + \beta_{7,i} oilprice_{t-1} + \beta_{8,i} eurnok_{t-1} + \beta_{9,i} fsr_{t,i} + \beta_{10,i} sent_{t-1} + \beta_{11,i}(pos_{t-1,i} \times sent_{t-1}) + \beta_{12,i}(neg_{t-1,i} \times sent_{t-1}) + \alpha_{1,i}\epsilon^2_{t-1,i} + u_{t,i}$$

The output from running the regression equation on the daily data is given in Table 25.

Appendix B.3.2. Monthly traded monetary value

Traded monetary value in month t for ticker i, $T_{t,i}$, was regressed using a first-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i in month $t-1$ and in interaction with the dummy variable sentiment index $sent_{t-1}$. A number of control variables were also added and those proving significant were kept. Formally:

$$T_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t-1,i} + \beta_{2,i} neu_{t-1,i} + \beta_{3,i} neg_{t-1,i} + \beta_{4,i}\sigma_{t-1} + \beta_{5,i} ro_{t-1} + \beta_{6,i} nibor_{t-1} + \beta_{7,i} oilprice_{t-1} + \beta_{8,i} eurnok_{t-1} + \beta_{9,i} fsr_{t,i} + \beta_{10,i} sent_{t-1} + \beta_{11,i}(pos_{t-1,i} \times sent_{t-1}) + \beta_{12,i}(neg_{t-1,i} \times sent_{t-1}) + \alpha_{1,i}\epsilon^2_{t-1,i} + u_{t,i}$$

The output from running the regression equation on the daily data is depicted in Table 26.

Appendix B.3.3. Average monhtly order size

Order size in month t for ticker i, $O_{t,i}$, was regressed using a first-order ARCH model with independent variables expressing the number of published positive, neutral and negative articles on ticker i in month $t-1$ and in interaction with the dummy variable sentiment index $sent_{t-1}$. Control variables were also added and those significant were kept. Formally:

$$O_{t,i} = \beta_{0,i} + \beta_{1,i} pos_{t-1,i} + \beta_{2,i} neu_{t-1,i} + \beta_{3,i} neg_{t-1,i} + \beta_{4,i}\sigma_{t-1} + \beta_{5,i} ro_{t-1} + \beta_{6,i} nibor_{t-1} + \beta_{7,i} oilprice_{t-1} + \beta_{8,i} eurnok_{t-1} + \beta_{9,i} fsr_{t,i} + \beta_{10,i} sent_{t-1} + \beta_{11,i}(pos_{t-1,i} \times sent_{t-1}) + \beta_{12,i}(neg_{t-1,i} \times sent_{t-1}) + \alpha_{1,i}\epsilon^2_{t-1,i} + u_{t,i}$$

The output from running the regression equation on the daily data is shown in Table 27.

Appendix C. Regression Output

In this appendix we present the results of the statistical regressions, which were performed in Stata®. The first three tables were constructed on an intraday basis, the following group of three tables were constructed on a weekly basis whilst the latter three were based on monthly data. Specifically, this appendix shows models trying to predict daily return (Table 19), daily traded monetary value (Table 20), daily order size (Table 21), average weekly return (Table 22), weekly traded monetary value (Table 23), average weekly order size (Table 24), monthly return (Table 25), monthly traded monetary value (Table 26), and, lastly, average monthly order size (Table 27).

	FUNCOM $R_{t,i}$	IOX $R_{t,i}$	NAUR $R_{t,i}$	NOR $R_{t,i}$	NSG $R_{t,i}$	RCL $R_{t,i}$	SDRL $R_{t,i}$	STL $R_{t,i}$	TEL $R_{t,i}$	YAR $R_{t,i}$
$pos_{t,i}$	3.3659***	1.5632***	2.0991***	1.1362**	2.0040***	0.3486	0.2053	0.1978**	0.4427***	0.6271***
	(4.62)	(3.78)	(4.94)	(2.63)	(5.03)	(1.71)	(1.34)	(3.20)	(5.00)	(5.60)
$neu_{t,i}$	3.3849***	3.1871***	2.0974***	1.0594***	2.5201***	0.9280***	0.3599***	0.1409**	0.1907**	0.2952***
	(9.11)	(7.34)	(6.14)	(3.92)	(10.14)	(6.92)	(4.19)	(2.99)	(2.80)	(3.73)
$neg_{t,i}$	-6.2955***	-0.8222	-8.6403***	-3.1221***	-2.5176***	-0.9557***	-0.6620***	-0.5130***	-0.4245**	-1.2064***
	(-12.82)	(-1.46)	(-14.86)	(-5.47)	(-7.64)	(-4.32)	(-3.87)	(-6.31)	(-2.79)	(-8.43)
$\sigma_{t,t} - 30$	13.4677	-9.4150	16.1475	37.7160***	-5.4486	37.5121***	24.7144**	8.5286	30.8094***	36.4509***
	(1.33)	(-0.80)	(1.37)	(3.46)	(-0.39)	(4.68)	(3.23)	(1.59)	(5.89)	(4.52)
$nibor_{t-1}$	-0.1228	0.0949	-0.0490	-0.0675	-0.0780	-0.2803***	-0.0239	0.0255	-0.1215***	-0.0682
	(-1.38)	(1.61)	(-0.65)	(-0.99)	(-0.96)	(-5.75)	(-0.65)	(0.79)	(-3.89)	(-1.57)
$fsr_{t,i}$	0.5085	-1.5743**	1.5138***	0.0802	-0.3950*	0.9696***	-0.0724	-0.2160**	0.1209	0.1630
	(0.82)	(-2.78)	(5.11)	(0.39)	(-2.37)	(8.26)	(-0.34)	(-2.76)	(1.49)	(1.43)
$pos_{t,i} \times sent_t$	0.2224	0.3795	-0.5551	-0.5504	0.4735	0.0604	0.0176	-0.0658	-0.0478	-0.1771
	(0.26)	(0.71)	(-1.06)	(-1.07)	(0.97)	(0.23)	(0.09)	(-0.75)	(-0.39)	(-1.07)
$neg_{t,i} \times sent_t$	0.7259	0.5075	7.7964***	1.0465	0.2094	-0.4310	0.1797	0.1927*	-0.0200	0.5088*
	(1.18)	(0.69)	(11.52)	(1.85)	(0.54)	(-1.11)	(0.83)	(2.08)	(-0.12)	(2.51)
$pos_{t,i} \times h_t$	-3.0839	2.9778***	-0.1067	1.0276	-3.1320	0.6099	0.1165	-0.2099	-0.1239	-0.2056
	(-1.66)	(4.16)	(-0.04)	(0.64)	(-1.92)	(1.57)	(0.29)	(-0.90)	(-0.58)	(-0.54)
$neg_{t,i} \times h_t$	3.1034*	-8.6385***	9.8841***	3.1636	4.8514*	-0.7049	0.1213	-0.1010	0.1979	0.3525
	(2.51)	(-6.94)	(4.01)	(1.47)	(2.36)	(-0.88)	(0.25)	(-0.52)	(0.87)	(0.75)
Constant	0.1056	-0.2383	-0.2017	-0.4553*	0.1445	0.2596	-0.1774	-0.0498	-0.0258	-0.2290
	(0.41)	(-1.29)	(-0.81)	(-2.10)	(0.58)	(1.69)	(-1.45)	(-0.44)	(-0.23)	(-1.64)
N	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03
χ^2	374.5565	249.9855	352.3939	165.4827	259.2921	224.0961	47.2843	97.9946	120.7480	137.0855
LL	-4.6e+03	-4.5e+03	-4.5e+03	-4.3e+03	-4.5e+03	-3.9e+03	-3.4e+03	-3.0e+03	-3.2e+03	-3.7e+03
AIC	9.3e+03	9.1e+03	9.1e+03	8.7e+03	9.0e+03	7.8e+03	6.9e+03	6.1e+03	6.3e+03	7.3e+03
$p - value$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 19. Regression output from predicting intraday return by ticker

	FUNCOM $T_{t,i}$	IOX $T_{t,i}$	NAUR $T_{t,i}$	NOR $T_{t,i}$	NSG $T_{t,i}$	RCL $T_{t,i}$	SDRL $T_{t,i}$	STL $T_{t,i}$	TEL $T_{t,i}$	YAR $T_{t,i}$
$pos_{t,i}$	20.0323***	3.7834	4.6564***	40.7443***	12.6167***	53.9095***	13.9085***	8.4201***	15.1477***	31.1281***
	(15.23)	(1.66)	(3.59)	(5.98)	(8.32)	(35.28)	(7.29)	(7.47)	(5.86)	(22.42)
$neu_{t,i}$	34.8209***	75.9457***	-0.1237	33.9509***	4.3970**	26.5914***	9.3199***	4.9912***	8.8298***	4.8585***
	(20.53)	(34.59)	(-0.06)	(8.48)	(3.05)	(11.73)	(6.41)	(6.07)	(5.22)	(4.45)
$neg_{t,i}$	2.5897	-0.6404	-37.1790***	27.2136***	23.6821***	33.5638***	30.5187***	11.3216***	6.5302*	13.7219***
	(1.13)	(-0.15)	(-20.37)	(3.58)	(9.43)	(6.61)	(10.46)	(7.96)	(2.01)	(5.65)
ro_{t-1}	26.8920**	207.6933***	-20.4265	113.2152	440.1529***	137.0299***	-96.9027	36.7027	-41.4894	-79.5569*
	(2.71)	(26.18)	(-1.77)	(1.28)	(19.37)	(3.71)	(-1.70)	(1.15)	(-0.84)	(-2.30)
$\sigma_{t,t-30}$	574.0920***	-2.6e+03***	-4.0e+02***	-7.4e+03***	-4.7e+03***	-2.0e+03***	230.5165	-2.3e+03***	-3.3e+03***	-3.3e+03***
	(17.85)	(-87.20)	(-13.21)	(-33.80)	(-100.16)	(-21.06)	(1.29)	(-20.59)	(-20.89)	(-22.18)
$nibor_{t-1}$	0.6436**	13.4478***	-6.1093***	60.8167***	67.6700***	3.3076***	13.2378***	35.9732***	37.0128***	53.0869***
	(3.06)	(73.82)	(-44.32)	(34.11)	(294.00)	(5.04)	(19.81)	(69.87)	(46.56)	(89.47)
$oilprice_{t-1}$	-0.2131***	0.8804***	1.0670***	-0.8402***	-0.6533***	-0.9390***	-0.5172***	-1.1990***	-1.0545***	-0.8108***
	(-23.34)	(46.56)	(85.08)	(-12.73)	(-29.90)	(-18.11)	(-6.76)	(-27.09)	(-16.54)	(-14.79)
$eurnok_{t-1}$	-8.5831***	43.1161***	53.7340***	9.1862*	15.1510***	-16.4606***	-3.4451	0.7569	4.9515	5.9969**
	(-18.51)	(41.73)	(122.51)	(2.00)	(11.90)	(-5.75)	(-0.95)	(0.37)	(1.58)	(2.66)
$fsr_{t,i}$	17.9922***	-10.0403	0.3667	19.6239***	11.1105***	70.0139***	6.4528**	15.0751***	21.1210***	7.5032***
	(8.24)	(-1.82)	(0.21)	(6.10)	(13.00)	(29.88)	(2.61)	(8.12)	(14.05)	(5.25)
$pos_{t,i} \times sent_t$	-14.7452***	3.1894	0.5638	-52.1015***	-6.0694**	-56.0387***	-4.0635	-8.3701***	-10.7593***	-18.5269***
	(-7.96)	(0.90)	(0.28)	(-4.81)	(-3.27)	(-22.94)	(-1.31)	(-5.65)	(-3.69)	(-8.84)
$neg_{t,i} \times sent_t$	-5.3609	-5.9798	43.3978***	14.6690	-21.9959***	-21.4623**	-21.7062***	-14.8216***	-4.7557	-10.1572**
	(-1.92)	(-0.90)	(24.10)	(1.50)	(-7.65)	(-3.02)	(-5.53)	(-8.98)	(-1.34)	(-2.98)
$pos_{t,i} \times h_t$	-1.4e+02***	17.8443	-14.0723	-16.6339	-7.4093	-33.7865***	-20.9031*	-1.9787	-19.7313**	-24.9773***
	(-24.86)	(0.67)	(-0.62)	(-0.33)	(-0.42)	(-5.10)	(-2.21)	(-0.63)	(-3.01)	(-4.87)
$neg_{t,i} \times h_t$	937.5206***	13.6020	5.9487	-14.6478	-36.1649	-17.2096	-39.2183**	-13.2673***	-9.9291	3.5474
	(27.24)	(0.23)	(0.42)	(-0.12)	(-1.96)	(-1.07)	(-3.16)	(-5.32)	(-1.81)	(0.38)
Constant	79.8843***	-4.1e+02***	-4.9e+02***	9.7354	-1.2e+02***	292.2029***	113.1986***	113.7570***	91.8452**	9.9896
	(18.54)	(-43.08)	(-110.83)	(0.23)	(-10.93)	(11.08)	(3.39)	(5.96)	(3.19)	(0.49)
	(9.23)	(14.99)	(11.84)	(37.04)	(11.61)	(15.21)	(19.49)	(12.76)	(22.43)	(19.55)
N	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03	1.6e+03
χ^2	9.4e+03	2.8e+04	3.5e+04	3.0e+03	2.1e+05	4.7e+03	1.1e+03	8.2e+03	3.6e+03	1.5e+04
LL	-8.4e+03	-9.1e+03	-7.6e+03	-9.5e+03	-8.8e+03	-8.8e+03	-8.3e+03	-8.0e+03	-8.2e+03	-8.1e+03
AIC	1.7e+04	1.8e+04	1.5e+04	1.9e+04	1.8e+04	1.8e+04	1.7e+04	1.6e+04	1.6e+04	1.6e+04
$p-value$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 20. Regression output from predicting intraday traded monetary value by ticker

	FUNCOM $O_{t,i}$	IOX $O_{t,i}$	NAUR $O_{t,i}$	NOR $O_{t,i}$	NSG $T_{t,i}$	RCL $O_{t,i}$	SDRL $O_{t,i}$	STL $O_{t,i}$	TEL $O_{t,i}$	YAR $O_{t,i}$
$pos_{t,i}$	5.7471	22.3678	-1.8745	16.6707**	12.6167***	0.0225	2.6608	4.7967***	6.6214***	6.0295***
	(0.73)	(1.20)	(-0.28)	(2.88)	(8.32)	(0.01)	(1.57)	(5.73)	(4.64)	(5.16)
$neu_{t,i}$	8.3803	20.2680	0.5705	9.9158***	4.3970**	2.2807	1.6625	0.6738	1.9491	4.3250***
	(1.41)	(0.85)	(0.10)	(4.93)	(3.05)	(1.53)	(1.75)	(1.00)	(1.88)	(5.56)
$neg_{t,i}$	14.8998	-7.2643	-14.0122	10.6217	23.6821***	-5.0074*	-1.6502	-1.8093	1.1048	0.3919
	(1.14)	(-0.19)	(-0.55)	(1.54)	(9.43)	(-2.49)	(-0.63)	(-1.52)	(0.58)	(0.21)
ro_{t-1}	268.7963***	203.2318	-1.6e+02*	291.6543***	440.1529***	-0.7638	-11.2511	1.0650	-4.3042	46.3914
	(4.02)	(0.65)	(-2.18)	(4.85)	(19.37)	(-0.02)	(-0.34)	(0.03)	(-0.12)	(1.28)
$\sigma_{t,t} - 30$	-63.0179	-6.7e+03***	2.9e+03***	-3.0e+03***	-4.7e+03***	-4.5e+03***	-2.8e+03***	-3.1e+03***	-2.8e+03***	-2.9e+03***
	(-0.22)	(-7.45)	(12.14)	(-10.04)	(-100.16)	(-23.39)	(-20.26)	(-22.43)	(-17.29)	(-17.16)
$nibor_{t-1}$	15.5494***	43.7836***	-1.4e+02***	35.3314***	67.6700***	60.2734***	36.0083***	58.0938***	23.0532***	38.8001***
	(4.86)	(5.17)	(-48.59)	(13.49)	(294.00)	(27.80)	(26.97)	(32.21)	(14.45)	(20.88)
$oilprice_{t-1}$	0.4235***	-0.7512**	4.4753***	-2.4184***	-0.6533***	-1.8506***	-1.6371***	-1.4655***	-0.8332***	-1.3037***
	(5.47)	(-2.90)	(74.05)	(-26.88)	(-29.90)	(-29.69)	(-40.54)	(-37.89)	(-16.79)	(-27.87)
$eurnok_{t-1}$	49.8920***	34.9349*	308.8977***	23.4172***	15.1510***	22.2640***	10.9667***	13.8761***	15.0470***	23.2595***
	(12.73)	(2.43)	(107.63)	(6.91)	(11.90)	(7.99)	(5.57)	(6.76)	(6.93)	(10.95)
$fsr_{t,i}$	9.8741	-3.2662	5.6883	2.0999	11.1105***	5.7737**	-1.8421	3.4395	2.6409	0.8082
	(0.88)	(-0.12)	(0.97)	(0.39)	(13.00)	(2.70)	(-0.57)	(1.65)	(1.08)	(0.51)
$pos_{t,i} \times sent_t$	-3.6050	-24.1448	-3.4970	-13.9691	-6.0694**	6.5874*	2.6030	-4.7143***	-3.6713*	-0.5560
	(-0.34)	(-0.67)	(-0.41)	(-1.88)	(-3.27)	(2.10)	(1.29)	(-4.19)	(-2.00)	(-0.34)
$neg_{t,i} \times sent_t$	-14.0705	-6.2018	4.0254	-15.4038	-21.9959***	9.8134*	3.2532	1.4332	-3.4902	-2.5742
	(-1.07)	(-0.12)	(0.16)	(-1.93)	(-7.65)	(2.37)	(1.11)	(1.00)	(-1.40)	(-1.08)
$pos_{t,i} \times h_t$	-9.4741	98.1687	-2.7e+02***	-27.6722	-7.4093	-4.1968	-9.3305**	-3.6525	-8.1233	-8.9100
	(-0.51)	(1.02)	(-7.62)	(-0.72)	(-0.42)	(-0.61)	(-3.04)	(-1.05)	(-1.23)	(-1.78)
$neg_{t,i} \times h_t$	72.7378***	5.8064	95.5577**	-20.6281	-36.1649	3.8563	-10.2163	-6.9833	-3.1824	0.1003
	(4.43)	(0.04)	(2.73)	(-0.24)	(-1.96)	(0.30)	(-1.87)	(-1.90)	(-0.40)	(0.01)
Constant	-3.9e+02***	-1.3e+02	-2.5e+03***	87.6431*	-1.2e+02***	9.5167	119.8665***	31.5234	39.5206	-19.2065
	(-10.28)	(-1.01)	(-87.18)	(2.54)	(-10.93)	(0.34)	(6.31)	(1.57)	(1.86)	(-0.89)
N	1.3e+03	1.2e+03	1.2e+03	1.2e+03	1.6e+03	1.3e+03	1.3e+03	1.3e+03	1.3e+03	1.3e+03
χ^2	329.9007	76.0480	1.1e+05	5.9e+03	2.1e+05	3.0e+03	3.7e+03	4.1e+03	1.0e+03	3.3e+03
LL	-6.5e+03	-7.1e+03	-6.8e+03	-6.8e+03	-8.8e+03	-6.0e+03	-5.7e+03	-5.7e+03	-5.6e+03	-5.6e+03
AIC	1.3e+04	1.4e+04	1.4e+04	1.4e+04	1.8e+04	1.2e+04	1.1e+04	1.1e+04	1.1e+04	1.1e+04
$p - value$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 21. Regression output from predicting intraday average order size by ticker

	FUNCOM $R_{t,i}$	IOX $R_{t,i}$	NAUR $R_{t,i}$	NOR $R_{t,i}$	NSG $R_{t,i}$	RCL $R_{t,i}$	SDRL $R_{t,i}$	STL $R_{t,i}$	TEL $R_{t,i}$	YAR $R_{t,i}$
$pos_{t-1,i}$	0.8044	-0.0237	1.1722	-0.4781	0.8591	0.5253	0.2112	0.1760	0.1808	-0.1578
	(0.87)	(-0.03)	(0.97)	(-0.67)	(1.34)	(1.49)	(1.17)	(1.56)	(1.21)	(-0.83)
$neu_{t-1,i}$	-0.1373	2.2730*	0.9983	-0.0019	-1.1300	-0.1045	0.0705	-0.1731	0.0583	0.2419
	(-0.13)	(2.24)	(0.76)	(-0.00)	(-1.87)	(-0.26)	(0.31)	(-1.94)	(0.31)	(1.36)
$neg_{t-1,i}$	-2.3694***	-3.3654**	-2.1204**	-1.2265	-0.3167	0.1494	-0.3851	0.0409	-0.0125	-0.1294
	(-3.39)	(-3.27)	(-2.72)	(-1.30)	(-0.74)	(0.22)	(-1.13)	(0.36)	(-0.06)	(-0.54)
$\sigma_{t,t-4}$	24.1725	-42.5960	134.4406*	47.6645	-1.2e+02*	-25.5146	142.0690***	7.9099	19.6391	120.0991**
	(0.38)	(-0.81)	(2.03)	(0.88)	(-2.12)	(-0.64)	(3.84)	(0.30)	(0.72)	(3.16)
$nibor_{t-1}$	-0.6259	-0.0106	-0.8223	-0.2038	0.2556	-0.4549*	-0.6229***	0.0060	-0.1071	-0.2328
	(-1.62)	(-0.04)	(-1.78)	(-0.57)	(0.72)	(-2.15)	(-4.49)	(0.05)	(-0.81)	(-1.07)
Constant	1.7130	0.2054	-0.0916	-0.0229	0.5939	1.3465	0.3244	0.0122	-0.1271	-0.6310
	(1.36)	(0.22)	(-0.07)	(-0.02)	(0.51)	(1.46)	(0.49)	(0.02)	(-0.22)	(-0.81)
N	315.0000	312.0000	315.0000	319.0000	316.0000	319.0000	321.0000	324.0000	324.0000	323.0000
χ^2	18.8885	14.5690	12.4391	5.2852	10.9583	12.5531	30.9294	4.8468	5.5475	12.2767
LL	-1.1e+03	-1.1e+03	-1.1e+03	-1.1e+03	-1.1e+03	-1.0e+03	-9.3e+02	-8.6e+02	-8.9e+02	-1.0e+03
AIC	2.3e+03	2.1e+03	2.3e+03	2.2e+03	2.2e+03	2.1e+03	1.9e+03	1.7e+03	1.8e+03	2.0e+03
$p-value$	0.0020	0.0124	0.0292	0.3821	0.0522	0.0279	0.0000	0.4349	0.3528	0.0312

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 22. Regression output from predicting weekly return by ticker

	FUNCOM $T_{t,i}$	IOX $T_{t,i}$	NAUR $T_{t,i}$	NOR $T_{t,i}$	NSG $T_{t,i}$	RCL $T_{t,i}$	SDRL $T_{t,i}$	STL $T_{t,i}$	TEL $T_{t,i}$	YAR $T_{t,i}$
$pos_{t-1,i}$	52.7348***	12.3646***	-2.9153	-6.4185	-19.9362***	28.4601***	4.1614	2.3717*	3.4930	5.2322***
	(12.29)	(4.71)	(-1.12)	(-0.98)	(-9.01)	(12.47)	(1.76)	(2.24)	(1.64)	(3.39)
$neu_{t-1,i}$	-3.4932	4.5763	2.3068	-7.2508	2.9200	-2.2753	-1.4300	1.8518**	2.1118	1.4181
	(-0.77)	(0.71)	(0.63)	(-1.77)	(1.36)	(-0.85)	(-0.72)	(3.18)	(1.31)	(1.17)
$neg_{t-1,i}$	46.8081***	-1.3649	-11.5300*	32.6582***	7.2752**	-11.8630**	8.5165*	0.6985	0.1298	-1.8840
	(8.66)	(-0.36)	(-2.08)	(5.56)	(2.76)	(-3.23)	(2.13)	(0.70)	(0.05)	(-1.18)
$\sigma_{t,t-4}$	-2.4e+03***	-7.5e+02**	-2.3e+03***	-6.9e+03***	-3.3e+03***	-1.2e+03**	-1.3e+03**	-2.2e+03***	-2.2e+03***	-1.6e+03***
	(-8.23)	(-3.27)	(-15.67)	(-12.50)	(-11.73)	(-2.93)	(-3.16)	(-9.79)	(-7.23)	(-5.71)
$nibor_{t-1}$	11.8489***	13.7737***	-11.1882***	28.8112***	90.2808***	0.9397	17.6957***	29.7878***	29.0413***	43.2074***
	(6.49)	(10.20)	(-19.21)	(9.34)	(44.59)	(0.47)	(9.90)	(24.09)	(24.58)	(55.84)
$oilprice_{t-1}$	1.0712***	-1.3075***	1.5774***	-2.5458***	0.7110***	-0.8338***	-0.9922***	-1.0806***	-0.6401***	-0.7266***
	(9.21)	(-14.38)	(29.46)	(-9.91)	(7.19)	(-4.98)	(-6.91)	(-8.83)	(-4.44)	(-6.91)
$eurnok_{t-1}$	45.8637***	-38.7628***	102.4005***	18.7660	20.3374***	-27.3150***	-10.8666	5.9312	10.6383	0.7395
	(8.89)	(-8.61)	(41.96)	(1.52)	(6.45)	(-4.38)	(-1.59)	(1.00)	(1.86)	(0.16)
$fsr_{t,i}$	-47.9284***	-10.8250	7.4221***	-2.1016	-3.5152*	-0.2292	1.4342	-1.9897	1.1379	-2.8092
	(-11.20)	(-1.74)	(6.50)	(-0.27)	(-2.53)	(-0.04)	(0.29)	(-0.87)	(0.60)	(-1.90)
$pos_{t-1,i} \times sent_{t-1}$	-62.6173***	-3.5779	3.3714	-7.4491	26.6927***	-20.2781***	1.7771	-2.2516	-4.1571	-6.8540***
	(-10.35)	(-0.66)	(1.12)	(-0.85)	(10.48)	(-6.15)	(0.58)	(-1.61)	(-1.65)	(-3.39)
$neg_{t-1,i} \times sent_{t-1}$	3.8428	3.3616	8.6490	-30.2601***	-1.1600	8.3185	-9.9663*	-2.4824	-3.5021	3.4446
	(0.70)	(0.42)	(1.54)	(-3.95)	(-0.43)	(1.16)	(-2.01)	(-1.61)	(-1.20)	(1.17)
Constant	-4.5e+02***	423.5837***	-8.8e+02***	216.9657	-3.6e+02***	372.5583***	233.1642***	77.3818	17.5098	52.9742
	(-8.98)	(10.60)	(-38.22)	(1.89)	(-11.92)	(6.44)	(3.73)	(1.38)	(0.31)	(1.19)
N	321.0000	318.0000	317.0000	322.0000	324.0000	324.0000	324.0000	324.0000	324.0000	324.0000
χ^2	639.3693	507.6741	4.5e+03	404.0912	3.7e+03	254.7415	184.9734	1.1e+03	924.3678	8.5e+03
LL	-1.8e+03	-1.7e+03	-1.6e+03	-1.9e+03	-1.8e+03	-1.8e+03	-1.7e+03	-1.6e+03	-1.6e+03	-1.6e+03
AIC	3.7e+03	3.5e+03	3.2e+03	3.7e+03	3.5e+03	3.5e+03	3.3e+03	3.2e+03	3.3e+03	3.3e+03
$p-value$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 23. Regression output from predicting weekly traded monetary value by ticker

	FUNCOM $O_{t,i}$	IOX $O_{t,i}$	NAUR $O_{t,i}$	NOR $O_{t,i}$	NSG $O_{t,i}$	RCL $O_{t,i}$	SDRL $O_{t,i}$	STL $O_{t,i}$	TEL $O_{t,i}$	YAR $O_{t,i}$
$pos_{t-1,i}$	33.8999*	71.1605***	-4.2255***	-2.2396	13.5167	6.8280	-1.9853	1.2557	0.6989	11.9147***
	(2.37)	(7.33)	(-4.76)	(-0.48)	(1.37)	(0.68)	(-0.31)	(0.48)	(0.14)	(4.29)
$neu_{t-1,i}$	12.6320	6.2897	0.6139	-14.1624***	-2.0426	-6.3354	3.3824	4.4127	-1.3322	4.0456
	(0.94)	(0.56)	(0.56)	(-3.37)	(-0.15)	(-0.68)	(0.56)	(1.85)	(-0.33)	(1.54)
$neg_{t-1,i}$	9.2949	32.2127*	0.7331	12.1899	5.0913	17.4383	6.3930	-0.7644	5.4445	-9.7168
	(0.58)	(2.46)	(0.28)	(1.43)	(0.43)	(0.89)	(0.68)	(-0.23)	(0.99)	(-1.67)
$\sigma_{t,t-4}$	-3.0e+02	-6.9e+03***	-9.7e+02***	-5.1e+02	-6.2e+03**	-4.3e+03*	-2.1e+03	-3.6e+03**	-2.6e+03	-1.5e+03
	(-0.21)	(-6.23)	(-8.26)	(-0.59)	(-2.67)	(-2.00)	(-1.27)	(-2.62)	(-1.77)	(-0.97)
$nibor_{t-1}$	4.5893	47.8589***	13.3680***	-10.0155	54.2921**	37.3140*	17.0242	33.3420*	14.1310	18.4035
	(0.31)	(3.96)	(10.12)	(-1.06)	(2.68)	(2.08)	(0.91)	(2.56)	(1.15)	(1.17)
$oilprice_{t-1}$	0.5147	-1.2682***	-0.3839***	-2.2108***	-1.2516**	-1.2909*	-1.4277***	-1.4051***	-1.0115**	-1.0809**
	(1.16)	(-3.58)	(-10.39)	(-10.63)	(-2.59)	(-2.45)	(-3.81)	(-4.46)	(-2.97)	(-2.90)
$eurnok_{t-1}$	21.4999	23.9489	0.8647	55.2219***	40.8784	18.9334	7.2350	11.0165	6.5202	25.8830
	(1.15)	(1.51)	(0.43)	(5.04)	(1.51)	(0.72)	(0.39)	(0.74)	(0.51)	(1.69)
$fsr_{t,i}$	17.0001	-12.4446	-0.4116	-0.2994	-2.9703	4.4685	-8.3952	2.1641	3.8262	-3.5935
	(0.93)	(-0.76)	(-0.46)	(-0.05)	(-0.46)	(0.36)	(-0.57)	(0.39)	(0.80)	(-0.50)
$pos_{t-1,i} \times sent_{t-1}$	-22.0481	-52.2267***	5.5975***	-4.6467	5.5645	-4.7234	8.9850	1.4576	6.5070	-11.1126*
	(-1.21)	(-3.75)	(5.36)	(-0.54)	(0.36)	(-0.43)	(1.38)	(0.36)	(1.26)	(-1.99)
$neg_{t-1,i} \times sent_{t-1}$	-13.4372	-32.0762*	-1.0995	-11.8011	-11.9436	-2.6541	-13.3269	-5.4566	-14.1176	6.1899
	(-0.87)	(-2.10)	(-0.42)	(-1.28)	(-0.69)	(-0.11)	(-1.13)	(-1.11)	(-1.60)	(0.49)
Constant	-1.4e+02	-18.7590	25.4889	-91.7704	-1.6e+02	26.5086	155.8495	95.1030	140.3840	-36.4488
	(-0.75)	(-0.12)	(1.38)	(-0.84)	(-0.66)	(0.11)	(0.86)	(0.67)	(1.07)	(-0.25)
N	260.0000	258.0000	253.0000	260.0000	261.0000	260.0000	261.0000	261.0000	261.0000	261.0000
χ^2	27.8925	146.8266	427.2534	610.0713	24.2537	37.2121	62.3648	67.1407	29.5856	133.2593
LL	-1.5e+03	-1.5e+03	-1.1e+03	-1.5e+03	-1.5e+03	-1.5e+03	-1.4e+03	-1.4e+03	-1.4e+03	-1.4e+03
AIC	3.1e+03	3.1e+03	2.2e+03	3.0e+03	3.1e+03	3.0e+03	2.9e+03	2.8e+03	2.8e+03	2.9e+03
$p-value$	0.0019	0.0000	0.0000	0.0000	0.0070	0.0001	0.0000	0.0000	0.0010	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 24. Regression output from predicting average weekly order size by ticker

	FUNCOM	IOX	NAUR	NOR	NSG	RCL	SDRL	STL	TEL	YAR
	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$	$R_{t,i}$
$pos_{t-1,i}$	-2.6554	1.3357	5.7786	-0.1821	-0.8396	-0.2829	-0.3604	0.3258	-0.0025	0.1445
	(-0.48)	(0.29)	(0.74)	(-0.10)	(-0.37)	(-0.15)	(-0.42)	(1.09)	(-0.00)	(0.28)
$neu_{t-1,i}$	3.1994	-3.9434	6.5632*	-0.7913	2.5139	-0.3149	0.2129	-0.2043	-0.2664	0.4788
	(0.67)	(-0.79)	(2.30)	(-0.25)	(1.23)	(-0.13)	(0.29)	(-1.11)	(-0.38)	(0.94)
$neg_{t-1,i}$	-4.5900	-4.5515	-9.9794	2.1172	-0.8559	-0.1453	0.0661	0.2757	1.2228*	-1.2055
	(-0.95)	(-0.64)	(-0.68)	(0.34)	(-0.35)	(-0.04)	(0.05)	(0.96)	(1.98)	(-1.90)
ro_{t-1}	-2.8503	-67.2229	-52.9106	-1.7593	8.6580	-6.3991	11.6162	-22.0542	-7.3293	-2.2939
	(-0.04)	(-1.37)	(-0.57)	(-0.04)	(0.16)	(-0.19)	(0.50)	(-1.80)	(-0.29)	(-0.12)
σ_{t-1}	-2.1e+02	-4.0e+02	-4.8e+02	-45.6904	-6.5e+02	-1.6e+02	-70.4679	-2.4e+02*	298.0855	138.3906
	(-0.25)	(-0.59)	(-0.52)	(-0.09)	(-1.08)	(-0.46)	(-0.23)	(-1.98)	(1.23)	(0.54)
$nibor_{t-1}$	-1.9702	-0.7597	0.3351	1.1675	0.2211	-2.7650	-1.1806	0.1328	-3.3240*	-0.7189
	(-0.42)	(-0.22)	(0.06)	(0.47)	(0.08)	(-1.03)	(-0.78)	(0.19)	(-2.42)	(-0.68)
$oilprice_{t-1}$	0.1513	0.0587	-0.1923	-0.2823	0.0090	-0.0820	-0.1293	-0.0578	-0.0235	-0.0740
	(0.58)	(0.23)	(-0.71)	(-1.29)	(0.04)	(-0.53)	(-1.65)	(-1.16)	(-0.22)	(-0.73)
$eurnok_{t-1}$	15.1444	6.2103	16.1146	-8.3930	13.7185	4.3105	0.5236	3.5070	0.6143	-0.5965
	(1.49)	(0.53)	(1.56)	(-0.74)	(1.58)	(0.58)	(0.12)	(1.42)	(0.14)	(-0.10)
$fsr_{t,i}$	10.4077	4.7143	4.0639	0.2268	-0.9164	-1.3518	0.6071	-0.5319	-0.2858	0.1021
	(1.43)	(0.70)	(1.00)	(0.11)	(-0.43)	(-0.65)	(0.34)	(-0.71)	(-0.22)	(0.13)
$senti_{t-1}$	0.0089	0.0101	0.0083	0.0059	0.0043	-0.0019	0.0057*	0.0065***	0.0056	0.0066*
	(1.07)	(1.63)	(0.75)	(1.39)	(0.56)	(-0.33)	(2.50)	(4.91)	(1.74)	(2.08)
$pos_{t-1,i} \times sent_{t-1}$	-6.7840	1.4133	-5.7178	-1.1011	0.4126	0.3672	-0.3119	-0.2389	0.1461	-1.5219*
	(-1.12)	(0.26)	(-0.75)	(-0.32)	(0.15)	(0.18)	(-0.30)	(-0.61)	(0.14)	(-2.00)
$neg_{t-1,i} \times sent_{t-1}$	4.9064	-3.7682	8.6751	-2.5766	-0.0939	-0.0594	-0.3015	0.0232	-0.9911	1.4637
	(0.88)	(-0.47)	(0.59)	(-0.41)	(-0.03)	(-0.02)	(-0.20)	(0.05)	(-0.75)	(1.34)
Constant	-1.1e+02	-31.0435	-1.1e+02	99.6311	-98.0481	-14.2140	24.0657	-14.7165	8.6819	25.0223
	(-1.13)	(-0.29)	(-1.10)	(0.99)	(-1.16)	(-0.22)	(0.60)	(-0.67)	(0.21)	(0.48)
N	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000
χ^2	21.9066	12.7530	21.4421	15.0555	7.5843	10.4009	32.1735	39.4389	47.0947	36.9752
LL	-3.3e+02	-3.3e+02	-3.4e+02	-3.2e+02	-3.2e+02	-3.0e+02	-2.7e+02	-2.2e+02	-2.6e+02	-2.8e+02
AIC	693.7515	683.4827	715.3624	664.1872	664.8644	630.7163	561.9488	472.0305	548.4244	582.7721
$p-value$	0.0386	0.3872	0.0443	0.2384	0.8167	0.5808	0.0013	0.0001	0.0000	0.0002

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 25. Regression output from predicting monthly return by ticker

	FUNCOM $T_{t,i}$	IOX $T_{t,i}$	NAUR $T_{t,i}$	NOR $T_{t,i}$	NSG $T_{t,i}$	RCL $T_{t,i}$	SDRL $T_{t,i}$	STL $T_{t,i}$	TEL $T_{t,i}$	YAR $T_{t,i}$
$pos_{t-1,i}$	-8.9459*	6.2123**	-22.3254	1.2384	0.7256	1.2495	2.4311	-0.9917	2.0665	1.0775*
	(-2.50)	(3.14)	(-0.13)	(0.55)	(0.10)	(1.20)	(1.01)	(-0.80)	(0.93)	(2.45)
$neu_{t-1,i}$	36.0055***	-5.5494*	4.2181	3.7563	-6.8918	8.2316***	-2.5920	0.6875	-0.0249	3.1144***
	(8.82)	(-2.40)	(0.07)	(1.27)	(-1.03)	(7.91)	(-1.25)	(0.72)	(-0.02)	(9.97)
$neg_{t-1,i}$	-33.7612***	1.0221	-19.5547	10.4541	4.1340	12.8053***	1.7023	1.1460	-1.7326	-0.6178
	(-8.12)	(0.31)	(-0.06)	(1.40)	(0.85)	(4.56)	(0.38)	(0.90)	(-0.92)	(-1.12)
ro_{t-1}	5.1906	114.6337***	-2.7e+02	158.6557*	374.3600**	166.8862***	54.7901	172.1673**	34.9906	268.5035***
	(0.12)	(5.07)	(-0.15)	(2.13)	(3.22)	(5.03)	(0.86)	(2.63)	(0.58)	(9.10)
σ_{t-1}	-5.1e+03***	-8.2e+02**	-5.3e+03	-6.1e+03***	-6.0e+03**	1.9e+03***	-1.9e+03*	-3.3e+03***	-3.7e+03***	-3.6e+03***
	(-10.82)	(-2.80)	(-0.24)	(-4.03)	(-3.14)	(4.13)	(-2.44)	(-5.30)	(-6.71)	(-10.12)
$nibor_{t-1}$	34.6601***	10.5908***	-59.9815	20.7386***	75.4829***	-6.7965***	17.7868***	34.6427***	27.3035***	48.9888***
	(11.87)	(5.01)	(-0.59)	(3.58)	(9.80)	(-3.30)	(5.56)	(10.89)	(7.50)	(28.80)
$oilprice_{t-1}$	1.7735***	-0.0425	5.9671	-1.4624***	-1.5180*	0.0427	-0.8065***	-1.0223***	-1.0126***	-1.0267***
	(9.71)	(-0.60)	(0.70)	(-4.46)	(-2.32)	(0.31)	(-3.81)	(-4.66)	(-3.66)	(-11.39)
$eurnok_{t-1}$	110.3545***	7.7288*	312.5759	50.2810***	0.5434	-28.4353***	-14.9208	7.6369	4.5798	3.1254
	(13.67)	(2.28)	(1.45)	(4.38)	(0.02)	(-6.41)	(-1.81)	(0.73)	(0.43)	(0.71)
$fsr_{t,i}$	19.1629**	-2.0037	36.5001	0.7830	-2.7656	-2.9895**	-2.4116	-3.1888	0.7400	-2.3848***
	(3.29)	(-0.91)	(0.47)	(0.21)	(-0.53)	(-2.75)	(-0.46)	(-0.88)	(0.27)	(-3.63)
$senti_{t-1}$	0.0228***	0.0050	0.0708	-0.0199	-0.0061	-0.0056	-0.0161*	0.0088	0.0243**	-0.0033
	(5.13)	(1.41)	(0.39)	(-1.72)	(-0.34)	(-1.05)	(-1.99)	(1.38)	(2.92)	(-0.94)
$pos_{t-1,i} \times sent_{t-1}$	-10.3175*	0.6235	3.3232	1.4796	-0.5938	0.6091	1.8790	3.1441	-1.9161	-1.8628***
	(-2.14)	(0.31)	(0.02)	(0.36)	(-0.06)	(0.55)	(0.67)	(1.72)	(-0.90)	(-4.94)
$neg_{t-1,i} \times sent_{t-1}$	30.9838***	6.8902*	23.3253	-15.7073*	5.7847	-19.3148***	-4.5247	-5.1811*	-0.9449	-0.0937
	(7.58)	(2.21)	(0.07)	(-2.19)	(0.99)	(-6.74)	(-1.03)	(-2.54)	(-0.36)	(-0.18)
Constant	-1.0e+03***	-55.1815*	-2.6e+03	-1.7e+02	88.6177	258.8254***	245.6386***	93.0634	166.4782	55.0350
	(-13.07)	(-1.98)	(-1.15)	(-1.70)	(0.42)	(6.52)	(3.60)	(0.92)	(1.73)	(1.48)
N	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000	73.0000
χ^2	1.1e+05	9.0e+03	11.6350	114.6249	315.1796	502.9643	81.3158	416.6146	128.2568	6.2e+04
LL	-4.3e+02	-3.6e+02	-5.3e+02	-3.9e+02	-4.0e+02	-3.6e+02	-3.5e+02	-3.4e+02	-3.4e+02	-3.3e+02
AIC	897.9058	753.5686	1.1e+03	815.7823	819.5235	746.6951	723.2068	699.4290	707.6798	685.6107
$p-value$	0.0000	0.0000	0.4754	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 26. Regression output from predicting monthly traded monetary value by ticker

	FUNCOM $O_{t,i}$	IOX $O_{t,i}$	NAUR $O_{t,i}$	NOR $O_{t,i}$	NSG $O_{t,i}$	RCL $O_{t,i}$	SDRL $O_{t,i}$	STL $O_{t,i}$	TEL $O_{t,i}$	YAR $O_{t,i}$
$pos_{t-1,i}$	-12.2762	36.8034	-0.2664	8.8759	-8.5449	0.3842	-0.0473	-3.9082	2.6815	-0.7332
	(-0.60)	(1.02)	(-0.19)	(1.87)	(-1.41)	(0.18)	(-0.02)	(-1.92)	(1.33)	(-1.82)
$neu_{t-1,i}$	0.3116	-17.0859	12.6834***	-7.5308**	-2.8477	4.2129	-0.8961	0.1406	0.0371	-0.4420
	(0.02)	(-0.33)	(3.31)	(-2.69)	(-0.57)	(1.37)	(-0.36)	(0.16)	(0.02)	(-0.91)
$neg_{t-1,i}$	26.3359	-40.9995	-40.9781***	12.1463	14.7965	-12.6792	-0.3422	5.7551*	-0.2703	3.7082***
	(1.13)	(-1.18)	(-7.45)	(0.96)	(1.91)	(-1.88)	(-0.07)	(2.30)	(-0.12)	(16.98)
ro_{t-1}	424.2913	350.0354	65.9024***	73.6504	411.7436*	325.1414**	133.7542	61.0725	200.0302*	74.3520***
	(1.53)	(1.00)	(3.31)	(0.85)	(2.57)	(3.01)	(1.10)	(1.11)	(2.32)	(4.39)
σ_{t-1}	146.2318	-4.2e+03	-1.1e+03**	-3.0e+03**	-4.9e+03**	-2.5e+03*	-4.1e+03*	-4.6e+03***	-3.2e+03**	-3.4e+03***
	(0.04)	(-0.76)	(-3.06)	(-2.83)	(-2.73)	(-2.10)	(-2.27)	(-4.10)	(-2.65)	(-8.76)
$nibor_{t-1}$	19.2485	21.4937	-82.1452***	-4.1093	32.2957	16.5524	13.4992	39.9887***	24.5925*	33.9706***
	(0.49)	(0.32)	(-12.19)	(-0.32)	(1.40)	(1.34)	(0.65)	(3.55)	(2.13)	(7.22)
$oilprice_{t-1}$	0.1626	-0.7493	1.5172***	-1.9850***	-1.3335*	-1.2151***	-1.9669***	-2.0795***	-1.0961***	-1.9539***
	(0.18)	(-0.53)	(6.93)	(-6.49)	(-2.04)	(-4.92)	(-6.45)	(-11.07)	(-4.09)	(-13.07)
$eurnok_{t-1}$	-3.3702	8.0834	40.9055***	53.3062***	-2.3677	-12.9509	-11.4271	4.7779	-9.7035	-4.3008*
	(-0.08)	(0.13)	(5.32)	(3.43)	(-0.10)	(-1.60)	(-0.79)	(0.60)	(-0.71)	(-2.32)
$fsr_{t,i}$	16.0093	3.2711	12.8484***	6.0908	1.7915	-0.2726	18.4079*	-0.3613	1.1812	3.1142***
	(0.50)	(0.08)	(4.08)	(1.78)	(0.39)	(-0.12)	(2.41)	(-0.11)	(0.37)	(6.21)
$senti_{t-1}$	-0.0292	-0.0424	0.0136*	0.0122	-0.0669***	-0.0642***	-0.0226	-0.0155*	-0.0049	0.0019
	(-0.85)	(-0.60)	(2.15)	(0.86)	(-3.39)	(-4.33)	(-1.18)	(-1.98)	(-0.32)	(0.58)
$pos_{t-1,i} \times sent_{t-1}$	23.0086	-25.1832	-17.2949***	-3.0291	15.6710	-0.3465	-0.1541	6.2997**	-0.4407	0.5975
	(1.03)	(-0.55)	(-6.87)	(-0.62)	(1.66)	(-0.11)	(-0.03)	(2.81)	(-0.14)	(1.51)
$neg_{t-1,i} \times sent_{t-1}$	-19.5071	33.9340	38.9158***	-14.0852	-16.2851	9.0264	0.8135	-8.1990**	-0.5237	-2.6995***
	(-0.74)	(0.67)	(6.47)	(-1.23)	(-1.69)	(1.36)	(0.13)	(-3.02)	(-0.17)	(-12.81)
Constant	-6.8638	54.8839	-1.9e+02**	-92.0238	136.8020	228.6851**	367.1550**	198.2339**	239.0954	275.5198***
	(-0.02)	(0.09)	(-2.99)	(-0.70)	(0.54)	(2.92)	(2.63)	(2.65)	(1.96)	(15.82)
N	60.0000	60.0000	60.0000	60.0000	60.0000	60.0000	60.0000	60.0000	60.0000	60.0000
χ^2	12.0975	6.8185	1.1e+04	369.8441	63.0758	150.3746	108.5175	623.9095	51.3995	2.4e+04
LL	-3.4e+02	-3.6e+02	-3.6e+02	-3.1e+02	-3.1e+02	-2.9e+02	-2.9e+02	-2.8e+02	-2.8e+02	-2.6e+02
AIC	709.9611	751.0733	757.3739	659.8223	656.0885	603.3083	607.4899	588.7548	586.2803	553.1180
$p-value$	0.4379	0.8694	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 27. Regression output from predicting monthly average order size ticker

Appendix D. Summary of hypothesis tests

In this appendix we present a summary of the hypothesis tests performed on the regression output denoted in appendix C. The first three tables (table VIII, table IX, and table X), depict tests performed on intraday data. The following two tables represent tests performed on weekly data, (table XI and table XII). The final table listed in this appendix (table XIII) describes tests performed on monthly data. In the first column of each table a Z-test of variable significance with Bonferroni correction is performed (see Appendix E.E.1 for mathematical formulation). The second column exhibit the result of a Wald test of variable difference (see Appendix E.E.2 for mathematical formulation) whilst the last column shows the result of a One-sided Wald test of average variable difference (see Appendix E.E.4 for mathematical formulation). In addition to the mathematical details of the hypothesis tests performed in appendix E, Matlab implementations of these tests are presented in full code listings in appendix F.

	$H_0: \beta_i = 0, \forall i \in \{1,2,...,n\}$ $H_1: \exists \beta_i \neq 0, i \in \{1,2,...,n\}$	$H_0: \beta_i \neq \beta_j, \forall i,j \in \{1,2,...,n\}^2, i \neq j$ $H_1: \exists \beta_i \neq \beta_j, i,j \in \{1,2,...,n\}^2, i \neq j$	$H_0: \bar{\beta}_A = \bar{\beta}_B$ $H_1: \bar{\beta}_A < \bar{\beta}_B$
pos_t	✓***	✓***	✓*** †
neu_t	✓***	✓***	✓*** †
neg_t	✓***	✓***	✓*** ‡
$pos_t \times sb_t$			
$neg_t \times sb_t$	✓***	✓***	✓*** †
$pos_t \times h_t$	✓**	✓***	
$neg_t \times h_t$	✓***	✓***	✓** †

✓ Rejection of H_0
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
† Blue chip significantly smaller than white chip
‡ White chip significantly smaller than blue chip

Table VIII Summary of hypothesis tests when regressing intraday return

	$H_0: \beta_i = 0, \forall i \in \{1,2,...,n\}$ $H_1: \exists \beta_i \neq 0, i \in \{1,2,...,n\}$	$H_0: \beta_i \neq \beta_j, \forall i,j \in \{1,2,...,n\}^2, i \neq j$ $H_1: \exists \beta_i \neq \beta_j, i,j \in \{1,2,...,n\}^2, i \neq j$	$H_0: \bar{\beta}_A = \bar{\beta}_B$ $H_1: \bar{\beta}_A < \bar{\beta}_B$
pos_t	✓***	✓***	✓*** ‡
neu_t	✓***	✓***	✓*** †
neg_t	✓***	✓***	✓*** ‡
$pos_t \times sb_t$	✓***	✓***	✓* †
$neg_t \times sb_t$	✓***	✓***	✓*** †
$pos_t \times h_t$	✓***	✓***	
$neg_t \times h_t$	✓***	✓***	✓*** †

✓ Rejection of H_0
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
† Blue chip significantly smaller than white chip
‡ White chip significantly smaller than blue chip

Table IX Summary of hypothesis tests when regressing daily traded

	$H_0: \beta_i = 0, \forall i \in \{1,2,...,n\}$ $H_1: \exists \beta_i \neq 0, i \in \{1,2,...,n\}$	$H_0: \beta_i \neq \beta_j, \forall i,j \in \{1,2,...,n\}^2, i \neq j$ $H_1: \exists \beta_i \neq \beta_j, i,j \in \{1,2,...,n\}^2, i \neq j$	$H_0: \bar{\beta}_A = \bar{\beta}_B$ $H_1: \bar{\beta}_A < \bar{\beta}_B$
pos_t	✓***	✓***	
neu_t	✓***	✓***	
neg_t	✓***		
$pos_t \times sb_t$	✓***	✓**	
$neg_t \times sb_t$	✓***	✓***	
$pos_t \times h_t$	✓***	✓***	
$neg_t \times h_t$	✓***	✓***	

✓ Rejection of H_0
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
† Blue chip significantly smaller than white chip
‡ White chip significantly smaller than blue chip

Table X Summary of hypothesis tests when regressing daily order size

	$H_0: \beta_i = 0, \forall i \in \{1,2,...,n\}$ $H_1: \exists \beta_i \neq 0, i \in \{1,2,...,n\}$	$H_0: \beta_i \neq \beta_j, \forall i,j \in \{1,2,...,n\}^2, i \neq j$ $H_1: \exists \beta_i \neq \beta_j, i,j \in \{1,2,...,n\}^2, i \neq j$	$H_0: \bar{\beta}_A = \bar{\beta}_B$ $H_1: \bar{\beta}_A < \bar{\beta}_B$
pos_t	$\checkmark^{***}$	$\checkmark^{***}$	
neu_t	$\checkmark^{**}$		
neg_t	$\checkmark^{***}$	$\checkmark^{***}$	$\checkmark^{***}$ †
$pos_t \times sb_t$	$\checkmark^{***}$	$\checkmark^{***}$	
$neg_t \times sb_t$	$\checkmark^{***}$	$\checkmark^{**}$	

✓ Rejection of H_0
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
† Blue chip significantly smaller than white chip
‡ White chip significantly smaller than blue chip

Table XI Summary of hypothesis tests when regressing weekly traded

	$H_0: \beta_i = 0, \forall i \in \{1,2,...,n\}$ $H_1: \exists \beta_i \neq 0, i \in \{1,2,...,n\}$	$H_0: \beta_i \neq \beta_j, \forall i,j \in \{1,2,...,n\}^2, i \neq j$ $H_1: \exists \beta_i \neq \beta_j, i,j \in \{1,2,...,n\}^2, i \neq j$	$H_0: \bar{\beta}_A = \bar{\beta}_B$ $H_1: \bar{\beta}_A < \bar{\beta}_B$
pos_t	$\checkmark^{***}$	$\checkmark^{***}$	
neu_t	$\checkmark^{***}$	$\checkmark^{*}$	
neg_t			
$pos_t \times sb_t$	$\checkmark^{***}$	$\checkmark^{***}$	$\checkmark^{*}$ ‡
$neg_t \times sb_t$	$\checkmark^{*}$		

✓ Rejection of H_0
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
† Blue chip significantly smaller than white chip
‡ White chip significantly smaller than blue chip

Table XII Summary of hypothesis tests when regressing weekly order size

	$H_0: \beta_i = 0, \forall i \in \{1,2,...,n\}$ $H_1: \exists \beta_i \neq 0, i \in \{1,2,...,n\}$	$H_0: \beta_i \neq \beta_j, \forall i,j \in \{1,2,...,n\}^2, i \neq j$ $H_1: \exists \beta_i \neq \beta_j, i,j \in \{1,2,...,n\}^2, i \neq j$	$H_0: \bar{\beta}_A = \bar{\beta}_B$ $H_1: \bar{\beta}_A < \bar{\beta}_B$
pos_t	$\checkmark^{*}$	$\checkmark^{*}$	
neu_t	$\checkmark^{***}$	$\checkmark^{***}$	
neg_t	$\checkmark^{***}$	$\checkmark^{***}$	
$pos_t \times sb_t$	$\checkmark^{***}$	$\checkmark^{***}$	
$neg_t \times sb_t$	$\checkmark^{***}$	$\checkmark^{***}$	

✓ Rejection of H_0
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
† Blue chip significantly smaller than white chip
‡ White chip significantly smaller than blue chip

Table XIII Summary of hypothesis tests when regressing monthly traded

Appendix E. Mathematical details of hypothesis tests

Appendix E.1. Z-test of variable significance with Bonferroni correction

Given n regression models estimating the relationship between the same independent and dependent variable, a Z-test can be used to determine if the relationship between these variables is significant. We let the sets $B = \{\beta_1, \beta_2, ..., \beta_n\}$ and $\sigma = \{\sigma_{\beta_1}, \sigma_{\beta_2}, ..., \sigma_{\beta_n}\}$ denote the true regression coefficients and the standard deviations of the n regression models. These are estimated as $\hat{B} = \{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_n\}$ and $\hat{\sigma} = \{\hat{\sigma}_{\beta_1}, \hat{\sigma}_{\beta_2}, ..., \hat{\sigma}_{\beta_n}\}$. In order establish whether there is a relationship between the independent and dependent variable we test whether the null hypothesis, that all of the coefficients are equal to zero, is true versus the alternative hypothesis that at least one is different from zero. Formally:

$$H_0 : \beta_i = 0, \ \forall i \in \{1, 2, ..., n\}$$
$$H_1 : \exists \beta_i \neq 0, \ i \in \{1, 2, ..., n\}$$

Assuming $\hat{\beta}_i \sim \mathcal{N}\beta_i, \sigma^2_{\hat{\beta}_i}), \forall i \in B$, the Z statistic

$$Z_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}, \ \forall i \in \{1, 2, ..., n\}$$

will also be asymptotically normally distributed, $Z_i \sim \mathcal{N}(0, 1)$. Since we are performing multiple comparisons of β-coefficients from different regression models, Bonferroni correction must be used to maintain the correct familywise error rate (i.e. the probability of making a type I error). Bonferroni Correction states that rejecting all $p_i < \frac{\alpha}{n}$ will control that the familywise error rate will be below α (Abdi, 2007). Hence, a significance level of $\frac{\alpha}{n}$ must simply be employed in lieu of α when n such comparisons are made. The critical value is thus

$$Z_c = Z_{\frac{\alpha}{n}}$$

If

$$\exists \, |Z_i| > Z_c, \ i \in \{1, 2, ..., n\}$$

is satisfied we can reject H_0 and accept H_1. Conversely, if

$$|Z_i| \leq Z_c, \ \forall i \in \{1, 2, ..., n\}$$

we fail to reject H_0.

Appendix E.2. Wald test of variable difference

Given n regression models estimating the relationship between the same independent and dependent variable, a Wald test can be used to determine if the relationship between these variables is significantly different (Wooldridge, 2012). We let the vectors $\mathbf{B} = [\beta_1, \beta_2, ..., \beta_n]$ and $\boldsymbol{\sigma} = [\sigma_{\beta_1}, \sigma_{\beta_2}, ..., \sigma_{\beta_n}]$ denote the true regression coefficients and the standard deviations of the n regression models. These are estimated as $\hat{\mathbf{B}} = [\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_n]$ and $\hat{\boldsymbol{\sigma}} = [\hat{\sigma}_{\beta_1}, \hat{\sigma}_{\beta_2}, ..., \hat{\sigma}_{\beta_n}]$. In order to establish whether the relationships between the independent and dependent variables are different in the regression models we test whether the null hypothesis, that all of the coefficients are equal, is true versus the alternative hypothesis that at least two coefficients are non-equal. Formally:

$$H_0 : \beta_i \neq \beta_j, \ \forall i, j \in \{1, 2, ..., n\}^2, i \neq j$$
$$H_1 : \exists \beta_i \neq \beta_j, \ i, j \in \{1, 2, ..., n\}^2, i \neq j$$

Assuming $\mathbf{B}$ follows a multivariate normal distribution, $\hat{\mathbf{B}} \sim \ \mathcal{N}(\mathbf{B}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{B}$, and introducing $\boldsymbol{C}$[18], a $n-1 \times n$ contrast matrix of $\hat{\mathbf{B}}$ where each row sum equals 0, it can be shown that $\mathbf{C}\hat{\mathbf{B}} \sim \ \mathcal{N}(\mathbf{0}, \boldsymbol{C\Sigma C^T})$. Furthermore, it can be shown that the test statistic

$$W = \left(\boldsymbol{C}\hat{\boldsymbol{B}}\right)^T \left(\boldsymbol{C\Sigma C^T}\right)^{-1} \boldsymbol{C}\hat{\boldsymbol{B}}$$

[18] This contrast matrix can be chosen arbitrarily, of course in conformance with the stated requirements, since any contrast matrix, $\boldsymbol{C^*}$, can be written as the product of an $n \times n$ matrix, $\boldsymbol{A}$, and another contrast matrix, $\boldsymbol{C}$: $\boldsymbol{C^*} = \boldsymbol{AC}$. Multiplying both sides of $W = \left(\boldsymbol{C}\hat{\boldsymbol{B}}\right)^T \left(\boldsymbol{C\Sigma C^T}\right)^{-1} \boldsymbol{C}\hat{\boldsymbol{B}}$ with $\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}$ and $\boldsymbol{A}^{-1}\boldsymbol{A}$ yields $W = \left(\boldsymbol{C}\hat{\boldsymbol{B}}\right)^T \boldsymbol{A}^T(\boldsymbol{A}^T)^{-1} \left(\boldsymbol{C\Sigma C^T}\right)^{-1} (\boldsymbol{A}^{-1}\boldsymbol{A})\boldsymbol{C}\hat{\boldsymbol{B}}$. Applying elementary matrix operations on this equation gives $W = \left(\boldsymbol{AC}\hat{\boldsymbol{B}}\right)^T \left(\boldsymbol{AC\Sigma(AC)^T}\right)^{-1} \boldsymbol{AC}\hat{\boldsymbol{B}} = \left(\boldsymbol{C^*}\hat{\boldsymbol{B}}\right)^T \left(\boldsymbol{C^*\Sigma C^{*T}}\right)^{-1} \boldsymbol{C^*}\hat{\boldsymbol{B}}$ ■.

is χ^2 distributed with $n-1$ degrees of freedom. The critical value is this test statistic it thus

$$\chi_c^2 = \chi_{n-1,\alpha}^2$$

If

$$W > \chi_c^2$$

is satisfied we can reject H_0 and accept H_1. Conversely, if

$$W \leq \chi_c^2$$

we fail to reject H_0.

Appendix E.3. Two-sided Wald test of average variable difference

Given n regression models estimating the relationship between the same independent and dependent variable and a bipartite partition of these models, a Wald test can be used to determine if the relationship between these variables is significantly different for the two partitions (Wooldridge, 2012). We let the sets $B_A = \{\beta_{A,1}, \beta_{A,2}, ..., \beta_{A,k}\}$ and $\sigma_A = \{\sigma_{\beta_{A,1}}, \sigma_{\beta_{A,2}}, ..., \sigma_{\beta_{A,k}}\}$ denote the true regression coefficients and the standard deviations of the $k < n$ regression models ($m_A = \{1, 2, ..., k\}$) and $B_B = \{\beta_{B,1}, \beta_{B,2}, ..., \beta_{B,n-k}\}$ and $\sigma_B = \{\sigma_{\beta_{B,1}}, \sigma_{\beta_{B,2}}, ..., \sigma_{\beta_{B,n-k}}\}$ for the remaining $n-k$ models ($m_B = \{1, 2, ..., n-k\}$). These are estimated as $\hat{B}_A = \{\hat{\beta}_{A,1}, \hat{\beta}_{A,2}, ..., \hat{\beta}_{A,k}\}$, $\hat{\sigma}_A = \{\hat{\sigma}_{\beta_{A,1}}, \hat{\sigma}_{\beta_{A,2}}, ..., \hat{\sigma}_{\beta_{A,k}}\}$, $\hat{B}_B = \{\hat{\beta}_{B,1}, \hat{\beta}_{B,2}, ..., \hat{\beta}_{B,n-k}\}$, and $\hat{\sigma}_B = \{\hat{\sigma}_{\beta_{B,1}}, \hat{\sigma}_{\beta_{B,2}}, ..., \hat{\sigma}_{\beta_{B,n-k}}\}$, respectively. In order to establish whether the relationships between the independent and dependent variable are different for the two partitions we test whether the null hypothesis, that the average coefficient in both partitions are equal, is true versus the alternative hypothesis that the average is unequal for partition A than B. The average coefficient for the partitions is given by:

$$\bar{\beta}_A = \frac{1}{k} \sum_{i \in m_A} \beta_{A,i}$$
$$\bar{\beta}_B = \frac{1}{n-k} \sum_{i \in m_B} \beta_{B,i}$$

Formally the hypothesis being tested are:

$$H_0 : \bar{\beta}_A = \bar{\beta}_B$$
$$H_1 : \bar{\beta}_A \neq \bar{\beta}_B$$

Estimating $\bar{\beta}_A - \bar{\beta}_B$

$$\begin{aligned}\widehat{\bar{\beta}_A - \bar{\beta}_B} &= \hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B \\ &= \frac{1}{k}\sum_{i \in m_A} \hat{\beta}_{A,i} - \frac{1}{n-k}\sum_{i \in m_B} \hat{\beta}_{B,i}\end{aligned}$$

and $var(\bar{\beta}_A - \bar{\beta}_B)$

$$\begin{aligned}var(\bar{\beta}_A - \bar{\beta}_B) &= var(\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B) = var(\hat{\bar{\beta}}_A) + var(\hat{\bar{\beta}}_B) \\ &= \frac{1}{k^2}\sum_{i \in m_A} se(\hat{\beta}_{A,i})^2 + \frac{1}{(n-k)^2}\sum_{i \in m_B} se(\hat{\beta}_{B,i})^2\end{aligned}$$

and assuming $\hat{\beta}_{A,i} \sim \mathcal{N}(\beta_{A,i}, \sigma^2_{\hat{\beta}_{A,i}}), \forall i \in m_A$ and $\hat{\beta}_{B,i} \sim \mathcal{N}(\beta_{B,i}, \sigma^2_{\hat{\beta}_{B,i}}), \forall i \in m_B$, the χ^2 statistic[19] becomes

$$\begin{aligned}\chi^2 &= \left(\frac{\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B}{se(\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B)}\right)^2 \\ &= \left(\frac{\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B}{\sqrt{var(\hat{\bar{\beta}}_A) + var(\hat{\bar{\beta}}_B))}}\right)^2\end{aligned}$$

which will be asymptotically χ^2 distributed with 1 degree of freedom. The critical value is thus

$$\chi^2_c = \chi^2_{1,\alpha}$$

[19]Since the χ^2 distribution is the square to the normal distribution, a Z test statistic could just as easily have been used. The latter was done since using this squared distribution relieves explicit handling of the sign of the test statistic.

and if

$$\chi^2 > \chi_c^2$$

is satisfied we can reject H_0 and accept H_1. Conversely, if

$$\chi^2 \leq \chi_c^2$$

we fail to reject H_0.

Appendix E.4. One-sided Wald test of average variable difference

Given n regression models, like in Section E.E.3, estimating the relationship between the same independent and dependent variable and a bipartite partition of these models, a Wald test can be used to determine if the relationship between these variables is significantly different for the two partitions (Wooldridge, 2012). We let the sets $B_A = \{\beta_{A,1}, \beta_{A,2}, ..., \beta_{A,k}\}$ and $\sigma_A = \{\sigma_{\beta_{A,1}}, \sigma_{\beta_{A,2}}, ..., \sigma_{\beta_{A,k}}\}$ denote the true regression coefficients and the standard deviations of the $k < n$ regression models ($m_A = \{1, 2, ..., k\}$) and $B_B = \{\beta_{B,1}, \beta_{B,2}, ..., \beta_{B,n-k}\}$ and $\sigma_B = \{\sigma_{\beta_{B,1}}, \sigma_{\beta_{B,2}}, ..., \sigma_{\beta_{B,n-k}}\}$ for the remaining $n - k$ models ($m_B = \{1, 2, ..., n - k\}$). These are estimated as $\hat{B}_A = \{\hat{\beta}_{A,1}, \hat{\beta}_{A,2}, ..., \hat{\beta}_{A,k}\}$, $\hat{\sigma}_A = \{\hat{\sigma}_{\beta_{A,1}}, \hat{\sigma}_{\beta_{A,2}}, ..., \hat{\sigma}_{\beta_{A,k}}\}$, $\hat{B}_B = \{\hat{\beta}_{B,1}, \hat{\beta}_{B,2}, ..., \hat{\beta}_{B,n-k}\}$, and $\hat{\sigma}_B = \{\hat{\sigma}_{\beta_{B,1}}, \hat{\sigma}_{\beta_{B,2}}, ..., \hat{\sigma}_{\beta_{B,n-k}}\}$, respectively. In order establish whether the relationships between the independent and dependent variable are different for the two partitions we test whether the null hypothesis, that the average coefficient in both partitions are equal, is true versus the alternative hypothesis that the average is smaller for partition A than B. The average coefficient for the partitions is given by:

$$\bar{\beta}_A = \frac{1}{k} \sum_{i \in m_A} \beta_{A,i}$$
$$\bar{\beta}_B = \frac{1}{n-k} \sum_{i \in m_B} \beta_{B,i}$$

Formally the hypothesis being tested are:

$$H_0 : \bar{\beta}_A = \bar{\beta}_B$$
$$H_1 : \bar{\beta}_A < \bar{\beta}_B$$

Estimating $\bar{\beta}_A - \bar{\beta}_B$

$$\begin{aligned}\widehat{\bar{\beta}_A - \bar{\beta}_B} &= \hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B \\ &= \frac{1}{k}\sum_{i \in m_A} \hat{\beta}_{A,i} - \frac{1}{n-k}\sum_{i \in m_B} \hat{\beta}_{B,i}\end{aligned}$$

and $var(\bar{\beta}_A - \bar{\beta}_B)$

$$\begin{aligned}var(\bar{\beta}_A - \bar{\beta}_B) &= var(\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B) = var(\hat{\bar{\beta}}_A) + var(\hat{\bar{\beta}}_B) \\ &= \frac{1}{k^2}\sum_{i \in m_A} se(\hat{\beta}_{A,i})^2 + \frac{1}{(n-k)^2}\sum_{i \in m_B} se(\hat{\beta}_{B,i})^2\end{aligned}$$

and assuming $\hat{\beta}_{A,i} \sim \mathcal{N}(\beta_{A,i}, \sigma^2_{\beta_{A,i}}), \forall i \in m_A$ and $\hat{\beta}_{B,i} \sim \mathcal{N}(\beta_{B,i}, \sigma^2_{\beta_{B,i}}), \forall i \in m_B$, the Z statistic[20] becomes

$$\begin{aligned}Z &= \frac{\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B}{se(\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B)} \\ &= \frac{\hat{\bar{\beta}}_A - \hat{\bar{\beta}}_B}{\sqrt{var(\hat{\bar{\beta}}_A) + var(\hat{\bar{\beta}}_B))}}\end{aligned}$$

which will be asymptotically normally distributed, $Z \sim \mathcal{N}(0,1)$. The critical value is thus

$$Z_c = Z_\alpha$$

and if

$$Z < Z_c$$

is satisfied we can reject H_0 and accept H_1. Conversely, if

$$|Z| \geq Z_c$$

we fail to reject H_0.

[20]For the one-sided case, a Z test statistic, as oppose to a χ^2 statistic, must be used.

Appendix F. Code listings of hypothesis tests implemented in MatLab

Listing 1. zTestWithBonferroniCorrection.m: implementation of hypothesis test formulated in Appendix E.E.1

```
function [zBC, pValueBC] = zTestWithBonferroniCorrection(betaVector,
    seBetaVector)
    %{
        H_0: \beta_1 = \beta_2 = ... = \beta_n = 0
        H_1: At least one \beta_i != 0
    %}

    % Number of \betas considered in hypothesis test
    n = length(betaVector);

    % Compute z values
    zVector = betaVector ./ seBetaVector

    % Find z value with greatest absolute value
    z = zVector(find(abs(zVector)==max(abs(zVector))))

    % Calculation of smallest simple p-value
    pValue = 0.5 * (1 - normcdf(abs(z)))

    % Calculation of Bonferroni corrected p-value and z statistic
    pValueBC = pValue * n;
    zBC = norminv(pValueBC);
end
```

Listing 2. betaWaldTest.m: implementation of hypothesis test formulated in Appendix E.E.2

```
function [chiSq, pValue] = betaWaldTest(betaVector, seBetaVector)
    %{
        H0: \beta_1  = \beta_2 = ... = \beta_n = 0
        H1: \beta_1 != 0 || \beta_2 != 0 || ... || \beta_n != 0
    %}

    % Number of \betas considered in hypothesis test
    n = length(betaVector);

    % MLE of \hat \beta (weighted \beta average)
    weightVector = 1 ./ (seBetaVector .* seBetaVector);
    weightVector = weightVector / sum(weightVector);
    sigma = diag(seBetaVector .* seBetaVector);
    betaHatVector = sum(betaVector .* weightVector);

    % Difference between \beta_i and \hat \beta hat (weighted \beta
       average)
    delta = transpose(betaVector - betaHatVector);

    % Drop one row since delta only has n-1 linearly independent rows
    delta = delta(1:n-1);

    % Covariance matrix
    C = eye(n)- transpose(ones(1,n))* weightVector;
    C = C(1:n-1, 1:n);
    S = C * sigma * transpose(C);

    % Test statistic
    chiSq = transpose(delta) * ( S \ delta);

    % Computation of p-value using \chi squared distribution
    pValue = 1 - chi2cdf(chiSq, n-1);
```

```
end
```

Listing 3. betaAverageWaldTestTwoSided.m: implementation of hypothesis test formulated in Appendix E.E.3

```
function [chiSq, pValue] = betaAverageWaldTestTwoSided(betaVector1,
    seBetaVector1, betaVector2, seBetaVector2)
    %{
        H0: \bar \beta_1  = \bar \beta_2
        H1: \bar \beta_1 != \bar \beta_2
    %}

    % Number of \betas in each group considered in hypothesis test
    n = length(betaVector1);

    % \bar \beta for each group (average beta)
    beta1Hat = mean(betaVector1);
    beta2Hat = mean(betaVector2);

    % Variance of \bar \beta (average \beta)
    varBeta1Hat = sum(seBetaVector1 .* seBetaVector1) / (n*n);
    varBeta2Hat = sum(seBetaVector2 .* seBetaVector2) / (n*n);

    % Test statistic
    chiSq = ((beta1Hat - beta2Hat) / sqrt(varBeta1Hat + varBeta2Hat)) .^
        2;

    % Computation of p-value using \chi squared distribution
    pValue = 1 - chi2cdf(chiSq, 1);
end
```

Listing 4. betaAverageWaldTestOneSided.m: implementation of hypothesis test formulated in Appendix E.E.4

```
function [z, pValue] = betaAverageWaldTestOneSided(betaVector1,
    seBetaVector1, betaVector2, seBetaVector2)
    %{
        H0: \bar \beta_1 = \bar \beta_2
        H1: \bar \beta_1 < \bar \beta_2
    %}

    % Number of \betas in each group considered in hypothesis test
    n = length(betaVector1);

    % \bar \beta for each group (average beta)
    beta1Hat = mean(betaVector1);
    beta2Hat = mean(betaVector2);

    % Variance of \bar \beta (average \beta)
    varBeta1Hat = sum(seBetaVector1 .* seBetaVector1) / (n*n);
    varBeta2Hat = sum(seBetaVector2 .* seBetaVector2) / (n*n);

    % Test statistic
    z = (beta1Hat - beta2Hat) / sqrt(varBeta1Hat + varBeta2Hat);

    % Computation of p-value using normal distribution
    pValue = normcdf(z);
end
```

REFERENCES

Abdi, Herve, 2007, The bonferonni and šidák corrections for multiple comparisons, *Encyclopedia of measurement and statistics* 3, 103–107.

Anderson, John Robert, Ryszard S Michalski, Ryszard Stanisław Michalski, Jaime Guillermo Carbonell, and Tom Michael Mitchell, 1986, *Machine learning: An artificial intelligence approach*, volume 2 (Morgan Kaufmann).

Baker, Malcolm, and Jeffrey Wurgler, 2000, The equity share in new issues and aggregate stock returns, *The Journal of Finance* 55, 2219–2257.

Baker, Malcolm, and Jeffrey Wurgler, 2004, A catering theory of dividends, *The Journal of Finance* 59, 1125–1165.

Baker, Malcolm, and Jeffrey Wurgler, 2007, Investor sentiment in the stock market, *The Journal of Economic Perspectives* 21, 129–151.

Balahur, Alexandra, Ralf Steinberger, Mijail Alexandrov Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva, 2010, Sentiment analysis in the news., in *LREC*.

Barber, Brad, Terrance Odean, and Ning Zhu, 2006, Do noise traders move markets?, in *EFA 2006 Zurich Meetings Paper*.

Berry, Thomas D, and Keith M Howe, 1994, Public information arrival, *The Journal of Finance* 49, 1331–1346.

Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of econometrics* 31, 307–327.

Brown, Gregory W, and Michael T Cliff, 2004, Investor sentiment and the near-term stock market, *Journal of Empirical Finance* 11, 1–27.

Brown, Stephen J, William N Goetzmann, Takato Hiraki, Noriyoshi Shirishi, and Masahiro Watanabe, 2003, Investor sentiment in japanese and us daily mutual fund flows, Technical report, National Bureau of Economic Research.

Campbell, John Y, and W Andrew, 1997, *Lo, and A. Craig MacKinlay, 1997, The econometrics of financial markets* (Princeton University Press).

DeGennaro, Ramon P, and Ronald E Shrieves, 1997, Public information releases, private information arrival and volatility in the foreign exchange market, *Journal of Empirical Finance* 4, 295–315.

Ederington, Louis H, and Jae Ha Lee, 1993, How markets process information: News releases and volatility, *The Journal of Finance* 48, 1161–1191.

Edmans, Alex, Diego Garcia, and Øyvind Norli, 2007, Sports sentiment and stock returns, *The Journal of Finance* 62, 1967–1998.

Engle, Robert F, 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, *Econometrica: Journal of the Econometric Society* 987–1007.

Evans, David Kirk, Lun-Wei Ku, Yohei Seki, Hsin-Hsi Chen, and Noriko Kando, 2007, Opinion analysis across languages: An overview of and observations from the ntcir6 opinion analysis pilot task, in *Applications of Fuzzy Sets Theory*, 456–463 (Springer).

Fama, Eugene F, 1998, Market efficiency, long-term returns, and behavioral finance, *Journal of financial economics* 49, 283–306.

Fleming, Michael J, and Eli M Remolona, 1999, Price formation and liquidity in the us treasury market: The response to public information, *The Journal of Finance* 54, 1901–1915.

Frazzini, Andrea, and Owen A Lamont, 2008, Dumb money: Mutual fund flows and the cross-section of stock returns, *Journal of Financial Economics* 88, 299–322.

Gjerde, Øystein, and Frode Saettem, 1999, Causal relations among stock returns and macroeconomic variables in a small, open economy, *Journal of International Financial Markets, Institutions and Money* 9, 61–74.

Graham, John R, Jennifer L Koski, and Uri Loewenstein, 2006, Information flow and liquidity around anticipated and unanticipated dividend announcements*, *The Journal of Business* 79, 2301–2336.

Groß-Klußmann, Axel, and Nikolaus Hautsch, 2011, When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions, *Journal of Empirical Finance* 18, 321–340.

Hautsch, Nikolaus, and Dieter Hess, 2002, The processing of non-anticipated information in financial markets: Analyzing the impact of surprises in the employment report, *European Finance Review* 6, 133–161.

Jolliffe, IT, 1986, Principal component analysis, *Springer Series in Statistics, Berlin: Springer, 1986* 1.

Kalev, Petko S, Wai-Man Liu, Peter K Pham, and Elvis Jarnecic, 2004, Public information arrival and volatility of intraday stock returns, *Journal of Banking & Finance* 28, 1441–1467.

Kamstra, Mark J, Lisa A Kramer, and Maurice D Levi, 2003, Winter blues: A sad stock market cycle, *American Economic Review* 324–343.

Kim, Chan-Wung, and Jinwoo Park, 1994, Holiday effects and stock returns: further evidence, *Journal of Financial and Quantitative Analysis* 29, 145–157.

Kross, William, and Douglas A Schroeder, 1984, An empirical investigation of the effect of quarterly earnings announcement timing on stock returns, *Journal of Accounting Research* 153–176.

Landsman, Wayne R, and Edward L Maydew, 2002, Has the information content of quarterly earnings announcements declined in the past three decades?, *Journal of Accounting Research* 40, 797–808.

Lee, Charles, and Bhaskaran Swaminathan, 2000, Price momentum and trading volume, *The Journal of Finance* 55, 2017–2069.

Ljungqvist, Alexander, Vikram Nanda, and Rajdeep Singh, 2006, Hot markets, investor sentiment, and ipo pricing*, *The Journal of Business* 79, 1667–1702.

Lowenstein, Roger, 2000, *When genius failed: The rise and fall of Long-Term Capital Management* (Random House LLC).

Malatesta, Paul H, and Rex Thompson, 1985, Partially anticipated events: A model of stock price reactions with an application to corporate acquisitions, *Journal of Financial Economics* 14, 237–250.

Mian, G Mujtaba, and Srinivasan Sankaraguruswamy, 2008, Investor sentiment and stock market response to corporate news, in *European Finance Association Annual Meeting*.

Mitchell, Mark L, and J Harold Mulherin, 1994, The impact of public information on the stock market, *The Journal of Finance* 49, 923–950.

Mourad, Ahmed, and Kareem Darwish, 2013, Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs, *WASSA 2013* 55.

Neal, Robert, and Simon M Wheatley, 1998, Do measures of investor sentiment predict returns?, *Journal of Financial and Quantitative Analysis* 33, 523–547.

Njølstad, Pål-Christian Salvesen, Lars Smørås Høysæter, and Jon Atle Gulla, 2014a, Optimizing supervised sentiment lexicon acquisition: Selecting co-occurring terms to annotate for sentiment analysis of financial news, Submitted to Language Resources and Evaluation. Available at: `http://folk.ntnu.no/palchrnj/papers/paper_lre.pdf`, note.

Njølstad, Pål-Christian Salvesen, Lars Smørås Høysæter, Wei Wei, and Jon Atle Gulla, 2014b, Evaluating feature sets and classiers for sentiment analysis of financial news, To appear in the proceedings of the 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2014). Available at: `http://folk.ntnu.no/palchrnj/papers/paper_wic2014.pdf`, note.

Pontiff, Jeffrey, 2006, Costly arbitrage and the myth of idiosyncratic risk, *Journal of Accounting and Economics* 42, 35–52.

Robinson, George K, 1991, That blup is a good thing: The estimation of random effects, *Statistical Science* 15–32.

Rosen, Richard J, 2006, Merger momentum and investor sentiment: The stock market reaction to merger announcements*, *The Journal of Business* 79, 987–1017.

Russell, Stuart Jonathan, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards, 1995, *Artificial intelligence: a modern approach*, volume 2 (Prentice hall Englewood Cliffs).

Scheinkman, Jose A, and Wei Xiong, 2003, Overconfidence and speculative bubbles, *Journal of political Economy* 111, 1183–1220.

Seyhun, Hasan Nejat, 2000, *Investment intelligence from insider trading* (MIT press).

Shiller, Robert J, 2000, *Irrational exuberance* (Princeton University Press (Princeton, NJ)).

Shleifer, Andrei, and Lawrence H Summers, 1990, The noise trader approach to finance, *Journal of Economic perspectives* 4, 19–33.

Shleifer, Andrei, and Robert W Vishny, 1997, The limits of arbitrage, *The Journal of Finance* 52, 35–55.

Slotani, Mlnoru, 1964, Tolerance regions for a multivariate normal population, *Annals of the Institute of Statistical Mathematics* 16, 135–153.

Wei, Wei, and Jon Atle Gulla, 2010, Sentiment learning on product reviews via sentiment ontology tree, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 404–413, Association for Computational Linguistics.

Whaley, Robert E, 2000, The investor fear gauge, *The Journal of Portfolio Management* 26, 12–17.

Wooldridge, Jeffrey, 2012, *Introductory econometrics: A modern approach* (Cengage Learning).

(This page is intentionally left blank.)

Bibliography

Herve Abdi. The bonferonni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107, 2007.

Željko Agić, Nikola Ljubešić, and Marko Tadić. Towards sentiment analysis of financial texts in croatian. *Bull market*, 143(45):69, 2010.

John Robert Anderson, Ryszard Spencer Michalski, Ryszard Stanisław Michalski, Thomas Michael Mitchell, et al. *Machine learning: An artificial intelligence approach*, volume 2. Morgan Kaufmann, 1986.

Sarah L Andrews, Peenaki Dam, Damien Frennet, Summit Chaudhuri, Ricardo Rodriguez, Ashok Ganapam, Frank Schilder, and Jochen Lother Leldner. Methods and systems for generating composite index using social media sourced data and sentiment analysis, December 27 2011. US Patent App. 13/337,662.

Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152, 2007.

Alexandra Balahur, Ralf Steinberger, Mijail Alexandrov Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. In *LREC*, 2010.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*, 2013.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

Tove Bjørneset. Ordboksprosjektet nordlexin-n. *Tematiska bidrag*, page 35, 1999.

Tim Bollerslev. Glossary to arch (garch). Technical report, CREATES Research Paper 2008-49. doi: 10.2139/ssrn. 1263250, 2008.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

John Y Campbell and W Andrew. *Lo, and A. Craig MacKinlay, 1997, The econometrics of financial markets*. Princeton University Press, 1997.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233, 2006.

Gennaro Chierchia and Sally MacConnell-Ginet. *Meaning and grammar: An introduction to semantics*. MIT press, 2000.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

David Crystal. *The Cambridge encyclopedia of the English language*. Ernst Klett Sprachen, 2004.

Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *ACL*. Citeseer, 2007.

John P Dickinson and Kinandu Muragu. Market efficiency in developing countries: A case study of the nairobi stock exchange. *Journal of Business Finance & Accounting*, 21(1):133–150, 1994.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

Eugene F Fama. Efficient capital markets: A review of theory and empirical work*. *The journal of Finance*, 25(2):383–417, 1970.

Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

Nicky J. Ferguson, Jie Michael Guo, Nicky Herbert Y.T. Lam, and Dennis Philip. Media Sentiment and UK Stock Returns. Working paper available at `http://ideas.repec.org/p/dur/durham/2011_06.html`, January 2011.

Allister David John Furey. *Evolutionary robotics in high altitude wind energy applications*. PhD thesis, University of Sussex, 2012.

Øystein Gjerde and Frode Saettem. Causal relations among stock returns and macroeconomic variables in a small, open economy. *Journal of International Financial Markets, Institutions and Money*, 9(1):61–74, 1999.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7, 2007.

Benjamin Graham. *The Intelligent Investor, A Book of Practical Counsel*. Harper & Brothers Publishers, 1959.

Benjamin Graham, David Le Fevre Dodd, and Sidney Cottle. *Security analysis*. McGraw-Hill New York, 1934.

Axel Groß-Klußmann and Nikolaus Hautsch. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, 18(2):321–340, 2011.

Michael Alexander Kirkwood Halliday. *On language and linguistics*, volume 3. Continuum, 2006.

Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

Arne Thorvald Gierløff Hollum, Borre P Mosch, and Zoltán Szlávik. Economic sentiment: Text-based prediction of stock price movements with machine learning and wordnet. In *Recent Trends in Applied Artificial Intelligence*, pages 322–331. Springer, 2013.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Bernardo A Huberman and Lada A Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

IT Jolliffe. Principal component analysis. *Springer Series in Statistics, Berlin: Springer, 1986*, 1, 1986.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Yoosin Kim, Seung Ryul Jeong, and Imran Ghani. Text opinion mining to analyze news for stock market prediction. *Int. J. Advance. Soft Comput. Appl*, 6(1), 2014.

Mikhail K Kozlov, Sergei P Tarasov, and Leonid G Khachiyan. The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 20(5):223–228, 1980.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2012.

Paul Kroeger. *Analyzing grammar: An introduction*. Cambridge University Press, 2005.

Hui-Min Li and Ke-Cun Zhang. A decomposition algorithm for solving large-scale quadratic programming problems. *Applied mathematics and computation*, 173 (1):394–403, 2006.

Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

Nimrod Megiddo and Arie Tamir. Linear time algorithms for some separable quadratic programming problems. *Operations Research Letters*, 13(4):203–211, 1993.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45, 1997.

Ahmed Mourad and Kareem Darwish. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. *WASSA 2013*, page 55, 2013.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Sentiful: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F Smeaton. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 9–16. ACM, 2009.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

Kishore Papineni. Why inverse document frequency? In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

Mihir Parikh, Robert M Fabricant, and Ed Hicks. Sentiment analysis, January 19 2012. US Patent App. 13/353,982.

Viktor Pekar. Linguistic preprocessing for distributional classification of words. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 15–21. Association for Computational Linguistics, 2004.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. Learning sentiment lexicons in spanish. In *LREC*, pages 3077–3081, 2012.

Joël Plisson, Nada Lavrac, Dr Mladenić, et al. A rule based approach to word lemmatization. SiKDD 2004 at multiconference IS-2004, 2004.

Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer, 2006.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In *IJCAI*, volume 9, pages 1199–1204, 2009.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

John Ross Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

Robert P Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3): 458–464, 2012.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29, 2013.

František Simančík and Mark Lee. A ccg-based system for valence shifting for sentiment analysis. *Research in Computing Science*, 41:99–108, 2009.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

Songbo Tan, Yuefen Wang, and Xueqi Cheng. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 743–744. ACM, 2008.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

John S Uebersax. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101(1):140, 1987.

Matthias W Uhl. Reuters sentiment and stock returns. Technical report, KOF working papers//KOF Swiss Economic Institute, ETH Zurich, 2011.

Jonathan J Webster and Chunyu Kit. Tokenization as the initial phase in nlp. In *Proceedings of the 14th conference on Computational linguistics-Volume 4*, pages 1106–1110. Association for Computational Linguistics, 1992.

Wei Wei and Jon Atle Gulla. Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 404–413. Association for Computational Linguistics, 2010.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.

Jeffrey Wooldridge. *Introductory econometrics: A modern approach*. Cengage Learning, 2012.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE, 2003.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48, 2008.

Yuzheng Zhai, Arthur Hsu, and Saman K Halgamuge. Combining news and technical indicators in daily stock price trends prediction. In *Advances in Neural Networks–ISNN 2007*, pages 1087–1096. Springer, 2007.

Wenbin Zhang and Steven Skiena. Trading strategies to exploit blog and news sentiment. In *ICWSM*, 2010.